Inferring action structure and causal relationships in continuous sequences of human action

Daphna Buchsbaum
Buchsbaum@psych.utoronto.ca
University of Toronto, Department of Psychology, 100 St. George Street, 4th Floor, Sidney Smith Hall, Toronto, ON M5S 3G3, Canada

Thomas L. Griffiths, Dillon Plunkett and Alison Gopnik
tom_griffiths@berkeley.edu, dillonplunkett@berkeley.edu, gopnik@berkeley.edu
University of California, Berkeley, Department of Psychology, 3210 Tolman Hall # 1650, Berkeley CA 94720-1650, USA

Dare Baldwin
baldwin@uoregon.edu
Department of Psychology, 1227 University of Oregon, Eugene, OR 97403-1227, USA

Keywords: Statistical learning; Causal inference; Action segmentation; Rational analysis; Bayesian inference

Address for correspondence:
Daphna Buchsbaum
University of Toronto, Department of Psychology
100 St. George Street, 4th Floor
Sidney Smith Hall
Toronto, ON M5S 3G3, Canada
E-mail: Buchsbaum@psych.utoronto.ca **Phone:** +1 617 335 7525 **Fax:** +1 (617) 209 1068

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE
Inferring action structure and causal relationships in continuous sequences of human action

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California, Berkeley,Department of Psychology,3210 Tolman Hall,Berkeley,CA,94720-1650

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Cognitive Psychology, in press.

14. ABSTRACT

In the real world, causal variables do not come pre-identified or occur in isolation, but instead are embedded within a continuous temporal stream of events. A challenge faced by both human learners and machine learning algorithms is identifying subsequences that correspond to the appropriate variables for causal inference. A specific instance of this problem is action segmentation: dividing a sequence of observed behavior into meaningful actions, and determining which of those actions lead to effects in the world. Here we present a Bayesian analysis of how statistical and causal cues to segmentation should optimally be combined, as well as four experiments investigating human action segmentation and causal inference. We find that both people and our model are sensitive to statistical regularities and causal structure in continuous action, and are able to combine these sources of information in order to correctly infer both causal relationships and segmentation boundaries.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**98**	

Abstract

In the real world, causal variables do not come pre-identified or occur in isolation, but instead are embedded within a continuous temporal stream of events. A challenge faced by both human learners and machine learning algorithms is identifying subsequences that correspond to the appropriate variables for causal inference. A specific instance of this problem is action segmentation: dividing a sequence of observed behavior into meaningful actions, and determining which of those actions lead to effects in the world. Here we present a Bayesian analysis of how statistical and causal cues to segmentation should optimally be combined, as well as four experiments investigating human action segmentation and causal inference. We find that both people and our model are sensitive to statistical regularities and causal structure in continuous action, and are able to combine these sources of information in order to correctly infer both causal relationships and segmentation boundaries.

Inferring action structure and causal relationships in continuous sequences of human action

Introduction

Human social reasoning depends on understanding the relationship between actions, goals and outcomes. In order to understand the reasons behind others' behavior, we must be able to distinguish the unique actions we see others performing, and recognize the effects of these actions. Imagine watching someone coming home and opening their front door. To understand this simple scene, an observer needs to identify meaningful behaviors from within the continuous stream of motion they see, such as "exiting the car", "coming up the stairs" and "opening the door", which are themselves composed of smaller motion elements such as "standing up", "closing the car door", "taking a step", "reaching for the doorknob", and so on.

Determining which subsequences of motion go together hierarchically, and what outcomes they produce, is also an important instance of the more general problem of causal variable discovery (a similar problem – determining how spatially distributed observations should be encoded as variables – is discussed by Goodman, Mansinghka, & Tenenbaum, 2007). Consider the case of learning which actions are necessary to open a door by observing multiple performances, embedded in everyday scenes such as the one above. A learner might notice that people almost always grasp and then turn a doorknob before the door opens, but sometimes they pull a handle instead. They frequently insert a key into a lock and then turn it before trying the doorknob, but not always. Often, other actions precede the door opening as well – putting down groceries, fumbling around in a purse, ringing a doorbell, sliding a bolt – which of these are causally necessary and which are incidental? While this ambiguity can make causal learning more challenging, the presence of statistical variation can actually aid inference. Motions that do not consistently precede outcomes are less likely to be causally necessary. Motions that reliably appear together and, in fact, predict each other, are more likely to be coherent units, corresponding to intentional, goal-directed action.

There is now a large body of evidence suggesting that both infants and adults can use statistical patterns in spoken language to help solve the related problem of segmenting words from continuous speech (for a partial review, see Gómez & Gerken, 2000). Recently, Baldwin, Andersson, Saffran, and Meyer (2008) demonstrated that a similar sensitivity to statistical regularities in continuous action sequences may play an important role in action processing. However, a key difference between action segmentation and word segmentation is that intentional actions usually have effects in the world. In fact, many of the causal relationships we experience result from our own and others' actions, suggesting that understanding action may bootstrap learning about causation, and vice versa. Here we present a combination of experimental and computational approaches investigating how the ability to segment action and to infer its causal structure functions and develops.

We first introduce a Bayesian analysis of action segmentation and causal inference, which provides a rational analysis of how statistical and causal cues to segmentation should optimally be combined. Next, we present four experiments investigating how both people and our model use statistical and causal cues to action structure. Our first experiment demonstrates that people are able to segment statistically determined actions using only the co-occurrence patterns between motions. This experiment is also the first to demonstrate that the continuous boundary judgment measures used in event segmentation research align with the sequence discrimination measures traditionally used in the statistical segmentation literature. Our second experiment demonstrates that people experience these actions as coherent, meaningful, and most importantly, causal sequences. Our third experiment shows that people are able to extract the correct causal variables from within a longer action sequence, and that they find causal sequences to be more coherent and meaningful than other sequences with equivalent statistical structure. Our fourth experiment demonstrates that, when statistical and causal cues conflict, both sets of cues influence segmentation and causal inference, suggesting that action structure and causal structure are learned jointly and simultaneously, and demonstrates that these results are not accounted for by simpler heuristic models. We conclude by discussing our results in the context of broader work, as well as its implications for more generalized human statistical learning abilities.

Background

Many if not most of the causal outcomes we witness are the result of intentional human action. We must be able to distinguish the unique actions we see other people performing and recognize their effects in order to understand the reasons behind others' behavior, and in order to potentially bring about those effects ourselves. But before we can interpret actions, we first must identify meaningful behaviors within a continuous, dynamic stream of motion (Sharon & Wynn, 1998; Byrne, 1999). What cues do we use to do this? How might infants and young children begin to break into the behavior stream in order to identify intentional, goal-directed actions? Could the causal relationships between actions and their outcomes in the world help us understand action structure itself? How might we identify reaching, grasping, and turning and then group them into the action "opening the door"?

Prior research has shown that adults can segment common everyday behaviors into meaningful events (e.g., Newtson, Engquist, & Bois, 1977; Zacks, Tversky, & Iyer, 2001), and that they do so unconsciously and automatically (e.g., Zacks, Braver, et al., 2001; Speer, Swallow, & Zacks, 2003; Zacks, Speer, Swallow, & Maley, 2010). When asked to explicitly provide boundary judgments for videos of everyday intentional action, adults agree on the boundary locations (e.g., Newtson, 1973; Speer et al., 2003), and their boundary judgments correspond to the goals and intentions underlying the actor's behavior (e.g., Zacks, 2004; Zacks et al., 2010). People's boundary judgments are also sensitive to the hierarchical structure of human action – they are able to consistently segment actions at multiple levels of granularity (e.g., "reach, grasp, turn" vs "open door") (Zacks, Tversky, & Iyer, 2001; Zacks, Braver, et al., 2001; Hard, Tversky, & Lang, 2006; Newtson, 1973).

While a full understanding of human action requires knowledge about goals and intentions, even young infants demonstrate sensitivity to the boundaries between intentional action segments (Woodward & Sommerville, 2000; Baldwin, Baird, Saylor, & Clark, 2001; Saylor, Baldwin, Baird, & LaBounty, 2007; Hespos, Saylor, & Grossman, 2009) well before they are thought to have a fully developed theory of mind (e.g., Wellman & Liu, 2004). This suggests that there may also be low-level cues to intentional action structure available in human motion that infants might initially use to begin identifying meaningful actions, allowing them to bootstrap their way into a more fully-fledged understanding of human action. The existence of low-level cues to intentional action structure, such as changes in body pose and movement features, corresponding to the locations of adult boundary judgments is supported by a variety of recent work (Newtson et al., 1977; Zacks, 2004; Hard et al., 2006; Zacks, Kumar, Abrams, & Mehta, 2009; Meyer, DeCamp, Hard, & Baldwin, 2010; Hard, Recchia, & Tversky, 2011; Buchsbaum, Canini, & Griffiths, 2011).

Another potentially important source of information is statistical regularities in the action stream. It has frequently been suggested that the nature of intentional, goal-directed action means that actors' motions will often produce structured, predictable patterns (e.g., reaching often precedes grasping, turning the door knob often precedes opening the door), detectable even without explicit knowledge of the underlying goal structure that generated them (e.g., Newtson et al., 1977; Byrne, 1999; Baldwin et al., 2001; Reynolds, Zacks, & Braver, 2007). Therefore, one way that infants might be able to segment actions is by using statistical regularities in human motion. There is now a lot of evidence that both infants and adults use statistical patterns in spoken language to help solve the related problem of segmenting words from continuous speech (e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Aslin, Saffran, & Newport, 1998; Pelucchi, Hay, & Saffran, 2009). In these experiments, infants (and adults) listen to an artificial language constructed of made-up words, usually created from English syllables (e.g., dutaba, patubi, pidabu). The words are assembled into a continuous speech stream (e.g., dutabapatubipidabu.), with other potential segmentation cues such as intonation and pauses removed. In these experiments, as in many words in real languages, syllables within a word have higher transitional probabilities than syllables between words – you are more likely to hear ta followed by ba (as in dutaba) than to hear bi followed by pi (as in patubi pidabu). Both infants and adults are able to use these transitional probabilities in order to distinguish words in these artificial languages (dutaba, patubi, pidabu), from part-words – combinations of syllables that cross a word boundary (e.g., tabapa, tubipi), and from non-words, combinations of syllables that do not appear in the artificial language at all (e.g., dupapi, babibu).

Infants have also been shown to succeed at statistical language segmentation even when more naturalistic language stimuli are used (Lew-Williams, Pelucchi, & Saffran, 2011; Pelucchi et al., 2009). Likewise, infants have been shown to use conditional probabilities in the visual domain, to learn which components group together into visual objects (e.g., Fiser & Aslin, 2002) and in order to learn visual sequences over time (e.g., Kirkham, Slemmer, & Johnson, 2002).

Intriguingly, there is also evidence that children and adults can successfully map words learned through this type of segmentation to meanings (Mirman, Magnuson, Graf Estes, & Dixon, 2008; Graf Estes, Evans, Alibali, & Saffran, 2007; Hay, Pelucchi, Graf Estes, & Saffran, 2011) and, conversely, can use words they already know to help find boundaries and discover new words (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Similarly, a recent study shows that, in the visual domain, children use statistical patterns to infer the boundaries between objects, and then use that information to make further predictions about how objects will behave (Wu, Gopnik, Richardson, & Kirkham, 2011). So children do not just detect the statistics and then segment the streams accordingly. They actually treat those statistical units as if they were meaningful. Across domains, statistical learning can be viewed as a potentially powerful means by which infants can "break into" some more domain-specific inference task.

Recently, a similar sensitivity to statistical regularities in action sequences has been demonstrated in both adults (Baldwin et al., 2008; Meyer & Baldwin, 2011) and infants (Roseberry, Richie, Hirsh-Pasek, Golinkoff, & Shipley, 2011; Stahl, Romberg, Roseberry, Golinkoff, & Hirsh-Pasek, in press). Baldwin et al. (2008) extended the artificial language approach described above to the action domain, demonstrating that, just as people can recognize statistically coherent words from an artificial language, and distinguish them from non-words and part-words, they can also recognize artificial actions grouped only by statistical relationships, and can distinguish these sequences from non-actions (motions that never appeared together) and part-actions (motion sequences that cross an action boundary). Stahl et al. (in press) found similar results in 7-9 month old infants, showing that infants looked longer to part-action than to action sequences.

It is worth noting that, while the event segmentation experiments described earlier asked people to provide explicit boundary judgments for videos of everyday actions, interpretable in terms of goals and intentions, Baldwin et al. (2008) and Stahl et al. (in press) used sequences that, by design, were not inherently meaningful. In every day action, higher-level intentional structure and statistical cues are confounded (e.g., reach generally precedes grasp because of the overarching goal of retrieving an object). By making the combinations intentionally arbitrary, but statistically coherent, sensitivity to just statistical cues could be tested. As a result, this work used discrimination of isolated action and part-action sequences to measure segmentation performance. This approach is consistent with the previously described statistical word segmentation literature, on which these experiments were based. While it has been assumed that the ability to discriminate statistically coherent sequences from incoherent ones is the result of successful segmentation, the correspondence between discrimination measures and explicit boundary judgments has not yet been established.

Just as the ability to track statistical relationships appears to play an important role in word and action segmentation, recent work has demonstrated that both children and adults also rely on statistical relationships to make causal inferences, and can infer structured causal relationships from patterns of conditional probability between potential causes and their effects (e.g., Cheng, 1997; Gopnik et al., 2004; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011). In fact, though a variety of sources of information inform people's causal inferences (e.g., temporal information, mechanical knowledge, domain knowledge, social knowledge), patterns of conditional probability alone are sufficient for at least some causal inferences (e.g., Shanks, 1995; Cheng, 1997; Griffiths & Tenenbaum, 2005; Gopnik et al., 2004). In addition, Bayesian inference over causal graphical models has successfully been used to capture causal learning in both children (e.g., Gopnik et al., 2004; Sobel, Tenenbaum, & Gopnik, 2004; Schulz, Bonawitz, & Griffiths, 2007; Griffiths et al., 2011; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011) and adults (e.g., Glymour (1998); Rehder (2003); Tenenbaum and Griffiths (2001); Tenenbaum and Griffiths (2003); Griffiths and Tenenbaum (2005); Lu, Yuille, Liljeholm, Cheng, and Holyoak (2008); but see Bes, Sloman, Lucas, and Éric Raufaste (2012) for cases where this approach may not capture human behavior).

Further, problems of causal reasoning, language learning, and action parsing, and other tasks traditionally studied under the rubric of statistical learning, share a common problem of variable selection. That is, not only do learners have to discover the predictive patterns between variables in the domain, they have to discover what groupings of stimuli or features constitute the variables in the first place. In the case of language learning, infants not only have to discover which meanings correspond to which words, they have to discover the words themselves. Similarly, reasoning about causal relationships requires identifying which features group together into potential causes. However, previous work has generally assumed that the possible causes are known in advance. Figuring out how causal variables are identified from within a continuous sequence remains an important problem in this area.

In the same way that words have meanings, intentional actions usually lead to causal outcomes. In fact, as noted above, statistically coherent action sequences likely arise, at least in many cases, because the small-scale acts involved are causally linked in achieving a goal. This suggests that, just as identifying words assists in mapping them to meanings, segmenting human action may bootstrap learning about causation and vice versa. However, researchers have not yet explored whether action parsing and causal structure can be learned jointly.

Below, we introduce a Bayesian ideal observer model of action segmentation and causal inference. This model provides a computational level account of how an ideal learner, sensitive to statistical evidence in both domains, should identify meaningful and causal units of action from within a fluid stream. Just as Baldwin et al. (2008) modeled their statistical action experiments on previous experiments in statistical word segmentation, here we base our model on an ideal observer model of statistical word segmentation (Goldwater, Griffiths, & Johnson, 2009). There exists a natural analogy between the problems of segmenting speech and segmenting actions – both are dynamic sensory streams that pose similar challenges to the processor. We discuss this choice of model and describe the model in more detail in the following sections, as well as in Appendix A.

An Ideal Learner for Action Segmentation

We created an ideal learner model that jointly infers action segmentation and causal structure, using statistical regularities and temporal cues to causal relationships in an action stream. The role of this model is as an *ideal observer* (Geisler, 2003), meaning that it represents the best possible segmentation performance for the data, given a set of starting assumptions about how that data was generated (a *generative model*). Our "ideal learner" analysis is intended to be in the spirit of rational analysis (Anderson, 1990, 1991) and Marr's (1982) notion of a "computational level". We will focus on qualitative comparisons of people to the model, since our goal is to understand the capacities people have to use relevant statistical information and not to directly model the mechanisms by which that information is used.

Bayes' rule is often used as a way of modeling this type of ideal learner. Bayesian models work by assuming that a learner is evaluating a set of hypotheses about the state of the world, and has assigned a "prior" probability $P(h)$ to each hypothesis h in that set. Then, Bayes' rule indicates that after seeing data d, the learner should assign each hypothesis a "posterior" probability $P(h|d)$ proportional to $P(h)$ multiplied by the probability of observing d if h were true, $P(d|h)$. Bayes' rule is a principled way to combine inductive biases, represented as the prior distribution, with the evidence provided by data, using an explicit model of how the world generated that data. For instance, in the case of segmenting continuous speech into words, the observed data would be an unsegmented sequence of syllables, the prior an assumption that this sequence is actually composed of discrete, multi-syllable words with particular distributional properties, and a hypothesis would be a segmentation of that sequence into words (see Figure 1). Bayesian analysis gives us a principled way of determining what segmentation and causal inferences an ideal learner *should* make, given a set of assumptions about how the data was generated. It also forces us to clearly lay out those assumptions.

As discussed earlier, the computational problems of word segmentation and action segmentation are very similar. In both domains, a temporally transient, continuous stream must be divided into the discrete, hierarchically organized units that generated it. In both language and action, there are theoretically an infinite number of possible words or actions, but only a finite set that is ever actually heard or observed. The learner must therefore infer not only boundary locations, but also the word or action vocabulary that is being used to generate sentences or action sequences. Finally, in both domains, there is evidence that distributional cues, even in the absence of information about meaning or intentions, may be sufficient for performing at least some amount of segmentation.

Given these similarities, an ideal learner for statistical action segmentation should share many assumptions with one for statistical word segmentation. The Goldwater et al. (2009) model has already been empirically tested as a rational model of statistical word segmentation, providing a reasonable starting point for developing a similar model for action.

Goldwater et al. (2009) model the generative process for creating a speech stream as successively selecting words to add to the sequence one at a time. Here, we model sequences of action in exactly the same way, with actions composed of individual small-scale motions taking the place of words composed of syllables. In addition, we incorporated cause and effect information into the generative model, allowing some actions to be probabilistic causes. In the following sections, we describe these two pieces of our model – the model's assumptions about how the sequence of actions was generated, and the model's assumptions about how effects occurring during this sequence were caused – in more detail.

We will use this model to explore whether it is, in principle, possible to jointly infer action segmentation and causal structure using only the statistical co-occurrence information present in both domains. This is an interesting question, since it could be that other cues (e.g., intentional information, mechanistic causal cues) might be necessary to segment action sequences and to discover causally meaningful units of action. Similarly, we can contrast this model with models with even simpler assumptions, for instance those that perform segmentation and causal inference separately, and see whether the optimal segmentations under these models differ in meaningful ways.

Generative Model for Action Sequences

We first look at how a sequence of actions is generated. As noted earlier, our model for this process is exactly that used by Goldwater et al. (2009) to model sequences of words. This model has convenient properties for capturing our intuitions about how action sequences are generated, most importantly that any given action sequence is composed of a finite number of action types, but that there is a theoretically infinite space of possible

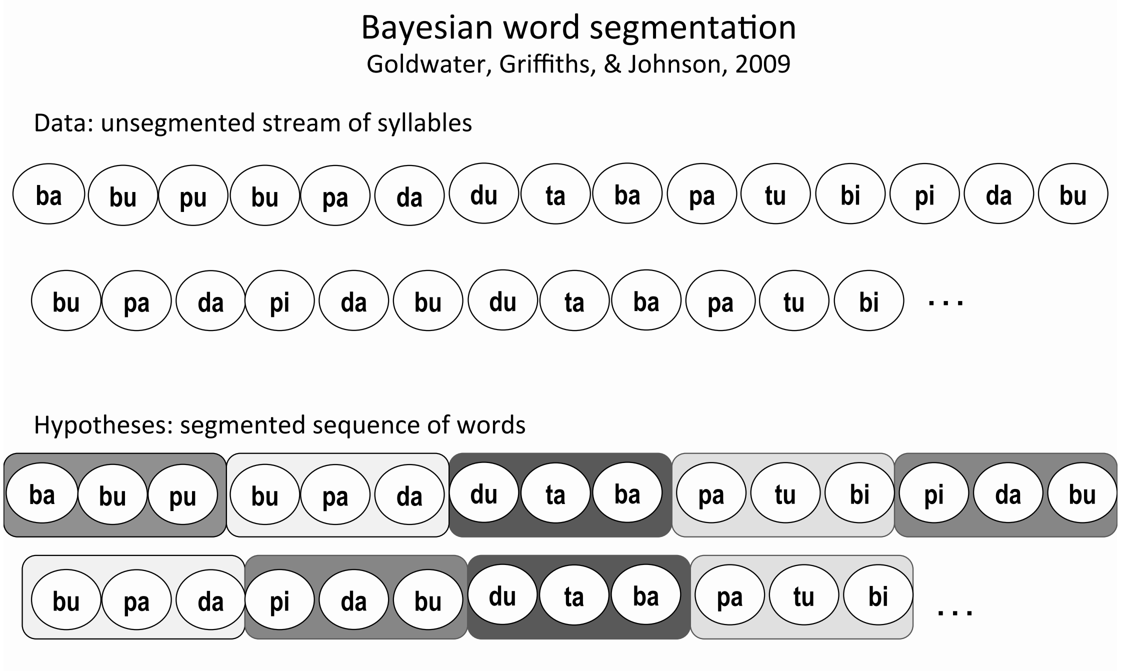

Bayesian word segmentation
Goldwater, Griffiths, & Johnson, 2009
Data: unsegmented stream of syllables
ba bu pu bu pa da du ta ba pa tu bi pi da bu
bu pa da pi da bu du ta ba pa tu bi . . .
Hypotheses: segmented sequence of words
ba bu pu
bu pa da
du ta ba
pa tu bi
pi da bu
bu pa da
pi da bu
du ta ba
pa tu bi
. . .

Figure 1. Example data and hypotheses for the Goldwater et al. (2009) model of word segmentation. Top: The data are a stream of unsegmented syllables. Bottom: A segmentation hypothesis, proposing a segmentation of this stream into words.

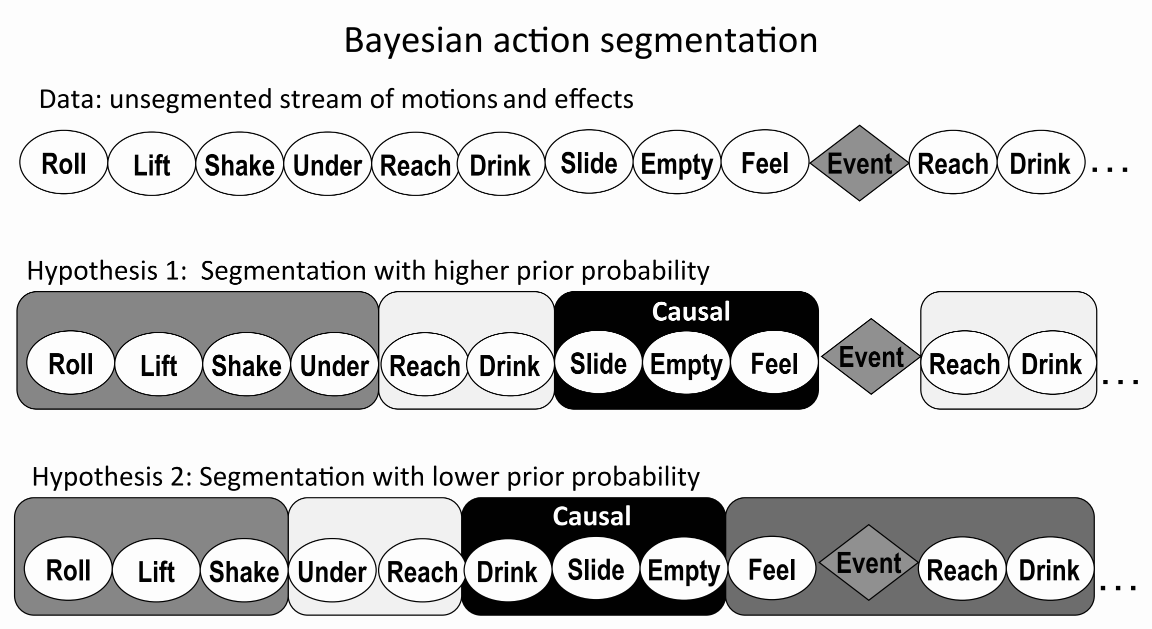

Bayesian action segmentation
Data: unsegmented stream of motions and effects
Roll Lift Shake Under Reach Drink Slide Empty Feel Event Reach Drink . . .
Hypothesis 1: Segmentation with higher prior probability
Roll Lift Shake Under
Reach Drink
Causal
Slide Empty Feel
Event
Reach Drink
. . .
Hypothesis 2: Segmentation with lower prior probability
Roll Lift Shake
Under Reach
Causal
Drink Slide Empty
Feel Event Reach Drink
. . .

Figure 2. Example data and hypotheses for our ideal observer action segmentation model. Top: The data are a stream of unsegmented motions. Bottom: Two segmentation hypotheses. Hypothesis 1, where an effect follows a causal action, is more likely than Hypothesis 2, where an effect occurs in the middle of an action, under our model.

actions.

Just as a sentence is composed of words, which are in turn composed of syllables, in our model an action sequence A is composed of actions a_i which are themselves composed of motion elements m_j. We assume a finite set of possible actions, and that complete actions are chosen one at a time from this set, and then added to the sequence. For each action a_i that is generated, we first decide whether it will be a novel action, or an action that has already occurred in the sequence. The probability of a previously occurring action is

$$p(a_i = \text{previously occurring action}|a_1, a_2, ..., a_{i-1}) = \frac{n}{n + \alpha_0} \tag{1}$$

and of a novel action is

$$p(a_i = \text{novel action}|a_1, a_2, ..., a_{i-1}) = \frac{\alpha_0}{n + \alpha_0} \tag{2}$$

where $a_1, a_2, ..., a_{i-1}$ are the actions already in the sequence, and n is the length of $a_1, a_2, ..., a_{i-1}$. So, the probability of performing a novel action depends on the *concentration parameter* α_0, and on the number of actions already performed. Intuitively, this means that the more actions you've performed, the less likely it is that the next action you perform will be one you've never done before. In this work we expect α_0 to be small, representing an expectation that the set of all possible actions is relatively small. This is consistent with the small vocabulary of actions used in previous statistical action segmentation experiments (Baldwin et al., 2008; Meyer & Baldwin, 2011; Stahl et al., in press), and in our own experiments described below.

If the next action in the sequence a_i is a previously occurring action, we must decide which one. The probability that the next action in the sequence will have a particular value $a_i = w$ is

$$p(a_i = w|a_1, a_2, ..., a_{i-1}) = \frac{n_w}{n} \tag{3}$$

where n_w is the number of times action w has already appeared. In other words, frequently occurring actions have a higher probability of being chosen again.

If the next action a_i is novel, then we need to generate its form. Actions are created by first choosing the action's length. We use a geometric distribution over length, which favors shorter actions

$$p(\text{length} = l) = p_{\#}(1 - p_{\#})^{l-1} \tag{4}$$

where l is the length of action a_i in motions and $p_{\#}$ is the probability of ending the action after each motion. In this work we expect $p_{\#}$ to be relatively large, which represents a bias towards finding smaller length actions, such as those used in previous statistical action segmentation experiments (Baldwin et al., 2008; Meyer & Baldwin, 2011; Stahl et al., in press), and in our own experiments described below.

To create the new action type, we randomly select motions one at a time, until the chosen length is reached. For simplicity, we assume that all motions are uniformly distributed.

Generative Model for Events

The action sequence A also contains non-action events e, which can occur between motions. We must now add a model of how these non-action events in our sequence are generated. The key assumption we would like to capture is that actions are more likely to be causal than other sequences of motion. An additional set of assumptions is motivated by previous research on causal inference, demonstrating that both children (Schulz & Sommerville, 2006; Griffiths et al., 2011; Buchsbaum, Gopnik, et al., 2011), and adults (Lu et al., 2008; Lucas & Griffiths, 2010; Yeung & Griffiths, 2011) strongly favor fewer, high probability causes.

To model the generation of causal effects, we use a probabilistic-OR model, commonly used to model human causal inference (Cheng, 1997; Griffiths & Tenenbaum, 2005), in which each potential causal variable has a probability π of being a cause, and each cause has an independent probability of generating the effect ω. There is also a

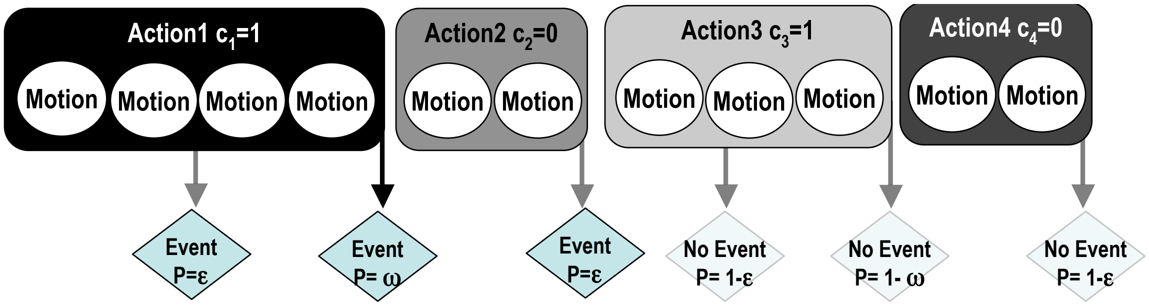


Figure 3. A theoretical action sequence depicting causal relationships in the model.

background probability of the effect occurring on its own (or due to unknown causes) ϵ.

In our model, the potential causal variables are the sequences that have been identified as actions. Every time a novel action type is generated, it is causal with probability π. We assume that π is small, which captures our assumption that relatively few actions are causes for a particular effect. If an action is a cause, then it is followed by an event with relatively high probability ω. We use a small value ϵ for the probability of an effect occurring after a non-causal sequence (in the middle of an action, or after a non-causal action, as shown in Figure 3). This captures our assumption that events are unlikely to follow non-causal sequences, and likely to occur after actions that are causes.

Inferring Segmentation and Causal Structure

An unsegmented action sequence consists of the motions m_j without any breaks between them, as well as the effects that occurred during this sequence. Given such a sequence, how do we find the boundaries between actions, and infer which actions are causal? A segmentation hypothesis h indicates whether there is an action boundary after each motion m_j, and whether each of the action types in the inferred vocabulary is causal. Each hypothesis also implicitly includes a proposed action vocabulary from which the sequence was composed. For a given segmentation hypothesis h, and unsegmented action sequence d, we use Bayes' rule $p(h|d) \propto p(d|h)p(h)$ to infer the posterior distribution $p(h|d)$. It is worth emphasizing that, unlike previous Bayesian models of causal inference (Griffiths & Tenenbaum, 2005, 2009), we do not start with a fixed, known set of potential

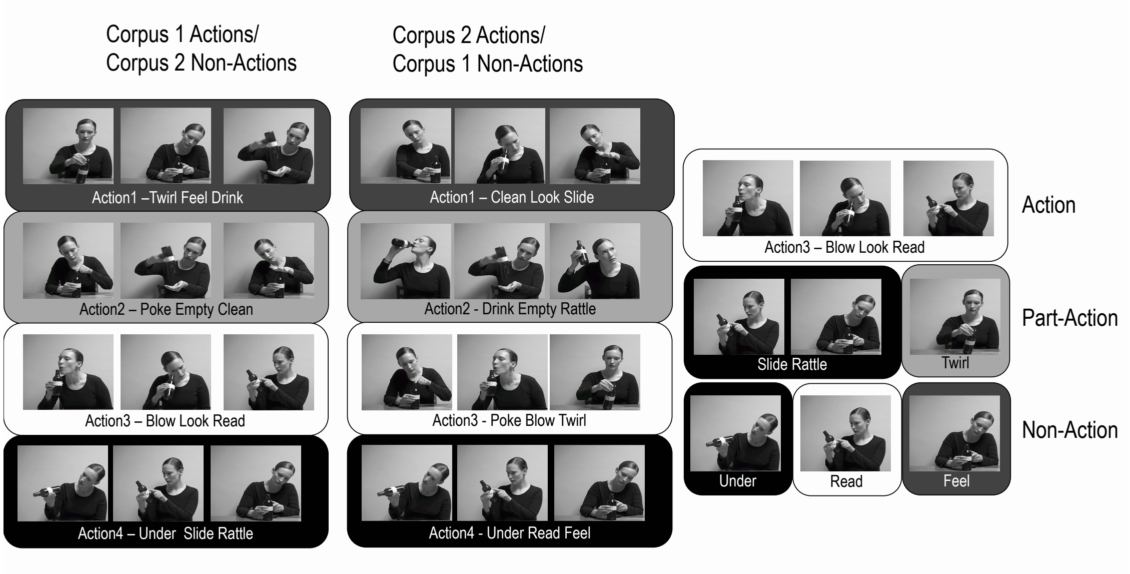

Corpus 1 Actions/
Corpus 2 Non-Actions
Corpus 2 Actions/
Corpus 1 Non-Actions
Action1 –Twirl Feel Drink
Action2 – Poke Empty Clean
Action3 – Blow Look Read
Action4 – Under Slide Rattle
Action1 – Clean Look Slide
Action2 - Drink Empty Rattle
Action3 - Poke Blow Twirl
Action4 - Under Read Feel
Action3 – Blow Look Read
Action
Slide Rattle
Twirl
Part-Action
Under
Read
Feel
Non-Action

Figure 4. Left: Four actions composed of three unique motions each were used to create the exposure corpus. Middle: A second corpus. The actions of corpus 1 are the non-actions of corpus 2. Right: Example Action, Part-Action and Non-Action for corpus 1.

causes. Instead, the sequences to be considered for causal inference must themselves be inferred. In our model, the sequences considered to be potential causes are precisely those proposed as actions.

We can estimate $p(h|d)$ using a standard Markov chain Monte Carlo method known as Gibbs sampling (Gilks, Richardson, & Spiegelhalter, 1996). The key property of a Gibbs sampler is that it allows us to sample segmentation hypotheses from the posterior distribution $p(h|d)$. Each sample corresponds to one possible segmentation of the corpus into actions, and includes not only the proposed boundaries but also implicitly includes a proposed action vocabulary from which the sequence was composed. For causal inference, each action type in the sample's vocabulary is also assigned a causal value. For all simulations described in this work, we ran three randomly seeded Gibbs samplers and averaged results from 10 samples from each sampler to estimate the posterior distributions and evaluate the model (for additional details see Appendix A as well as Goldwater et al. (2009)).

Predictions for Human Segmentation

Our model represents a number of assumptions about how fluid streams of motion are generated, including that i) motions are not produced randomly, but instead as coherent groupings, corresponding to meaningful units of action, which tend to appear repeatedly and predictably and ii) that some actions are causal, and are likely to generate non-action events.

The key intuition captured by this model is that action segmentation and causal structure are jointly learned, taking advantage of statistical evidence in both domains. The model represents our assumption that the same underlying process generates human actions and causal motion sequences, implicitly capturing that actions are being chosen intentionally, often to bring about causal outcomes. This means that in this model action segmentation and causal structure are inferred simultaneously and interdependently. Sequences of motion that correspond to known actions are considered more likely to be causes, and sequences of motion that appear to be causal (they predict outcomes in the world) are considered more likely to be actions. The inferred action boundaries help determine the inferred causal structure and vice versa. This corresponds to our hypothesis that people believe intentional actions and causal effects go hand in hand.

As an ideal observer model, our model's role is to provide a description of the best possible performance under our assumptions, rather than as a psychological process model meant to capture the precise mechanisms underlying human performance. Our goal in exploring this ideal model is to determine how statistical information should be used in identifying variables for causal learning. This then motivates our experiments, in which we explore qualitative predictions about the kinds of inferences people should make, if they are operating under similar assumptions, and compare those predictions to those made by alternative models.

First, if people are sensitive to the same types of statistical cues as our model then they should be able to segment sequences of action using these cues. This prediction is tested in Experiment 1. Experiment 1 also establishes a basic methodology for evaluating people's ability to identify actions, by demonstrating that explicit boundary judgments align with discrimination judgments. Second, if statistical action structure is in fact a cue to causal relationships then, like our model, people should think statistically grouped actions are more likely to be potential causes than other equivalent sequences. This prediction is tested in Experiment 2. Third, if people believe that causal sequences of motion are also likely to be actions, they should be able to identify and segment out causal sequences, and should find those sequences to be more meaningful and coherent than other sequences of motion with equivalent statistical regularities. This prediction is tested in Experiment 3. Finally, if action segmentation and causal relationships are truly jointly learned, then we should see cue combination and cue conflict effects emerge, as in other cases of joint perceptual inference (Ernst & Banks, 2002). This prediction is tested in Experiment 4.

Using Statistical Cues

In this first set of model simulations and experiments, we examined the boundary judgments produced by our model and by human participants, when presented with continuous sequences of motion generated from "artificial action grammars", similar to those used in previous action segmentation experiments (Baldwin et al., 2008; Meyer & Baldwin, 2011), and paralleling the designs used in the statistical word segmentation literature (e.g. Saffran, Aslin, & Newport, 1996; Saffran et al., 1997). Our subsequent simulations and experiments examine causal inferences from these same types of sequences.

Just as a sentence is composed of words, which are in turn composed of syllables, here an action sequence is composed of actions, which are themselves composed of small motion elements (SMEs). We created two exposure corpora, each assembled by adding "actions" to the sequence one at a time, selecting from four "actions" each made up of three distinct recognizable object-directed motions (see Table 1 for a description of the 12

Table 1
Small motion elements (SMEs) used in Experiments 1, 2, 3 and 4. After Meyer and Baldwin (2011).

SME	Description	Length (ms)
Empty	bottle is turned over as if to pour into open hand	1259
Clean	flat hand wipes top of bottle	1079
Under	bottom of bottle is examined	1139
Feel	index finger touches side of bottle in an up-and-down motion	1169
Blow	bottle is lifted to mouth and blown into	1049
Look	bottle is lifted to face and interior examined	1259
Drink	bottle is lifted and tipped into mouth as if drinking	1289
Twirl	bottle edge is lifted from table and spun around	959
Read	finger traces over label and bottle is lifted from table as if to read	1948
Rattle	bottle is lifted close to ear and shaken	1199
Slide	bottle is pushed forward on the table and returned	779
Poke	index finger is inserted and removed from top of bottle	869

motions used). Each action appears 90 times for a total of 360 actions and 1080 motions. A more detailed description of how the corpora were constructed is given in the methods section of Experiment 1a. The key feature of these corpora is that within an action the transitional probabilities between adjacent motion elements are higher (1.0 in all cases) than between actions (≈0.33). In this first set of experiments, no causal outcomes were added to the corpora. We first ran the model on the exposure corpora to examine its segmentation performance and inferred action vocabulary, and then looked at human performance on these same corpora.

Model Simulations

An abstract representation of each unsegmented corpus was used as input to the model, with a letter standing for each SME. For example, the sequence `blow, look, read, twirl, feel, drink, poke, empty, clean` would be represented as `BLRTFDPEC` (see Appendix C for examples of complete corpora). Our model has two free segmentation parameters: $p_{\#}$, which influences action length, and α_0, which influences the number of unique action types.[1]

We evaluated model results across a wide range of parameter values, comparing our results to the true segmentation, and calculated average precision and recall scores across samples, commonly used metrics in the natural language processing literature (e.g., Brent, 1999; Venkataraman, 2001), and which were also used by Goldwater et al. (2009). Precision (P) is the percent of all actions in the produced segmentation that are correct, while recall (R) is the percent of all actions in the true segmentation that were found. In other words, following Brent (1999)

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{5}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{6}$$

For example, for the true sequence `BLR TFD PEC` a segmentation hypothesis of `BLR TFDPEC` would have $P = 1$ because the one boundary found also appears in the true segmentation, and $R \approx 0.33$, because only one out of the three true boundaries was found. Meanwhile, the hypothesis `BLR TFD P E C` would have $P = 0.5$ because only two out of the four proposed boundaries appear in the true segmentation, and $R = 1$, because both of the correct boundary locations were found. The correct segmentation `BLR TFD PEC` would have perfect precision and recall, $P = 1, R = 1$.

We ran simulations for $\alpha_0 \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ and $p_{\#} \in \{0.5, 0.7, 0.90.95, 0.99\}$. Overall, segmentation performance was quite good across parameter values, with perfect boundary precision for all parameter values, meaning that regardless of the parameter settings, the model never identified boundaries in incorrect locations. Recall results were more variable, and are shown in Figure 5. In general,

[1] There is also one additional model parameter we did not mention, $p_{\$}$, the prior probability of the action sequence terminating. Since the action sequence ends only once, the effect of $p_{\$}$ on segmentation results is negligible – preliminary simulations varying the value of $p_{\$}$ confirmed that the exact value had little influence on the resulting segmentation. We therefore used a fixed value of $p_{\$} = 0.01$ for our simulations.

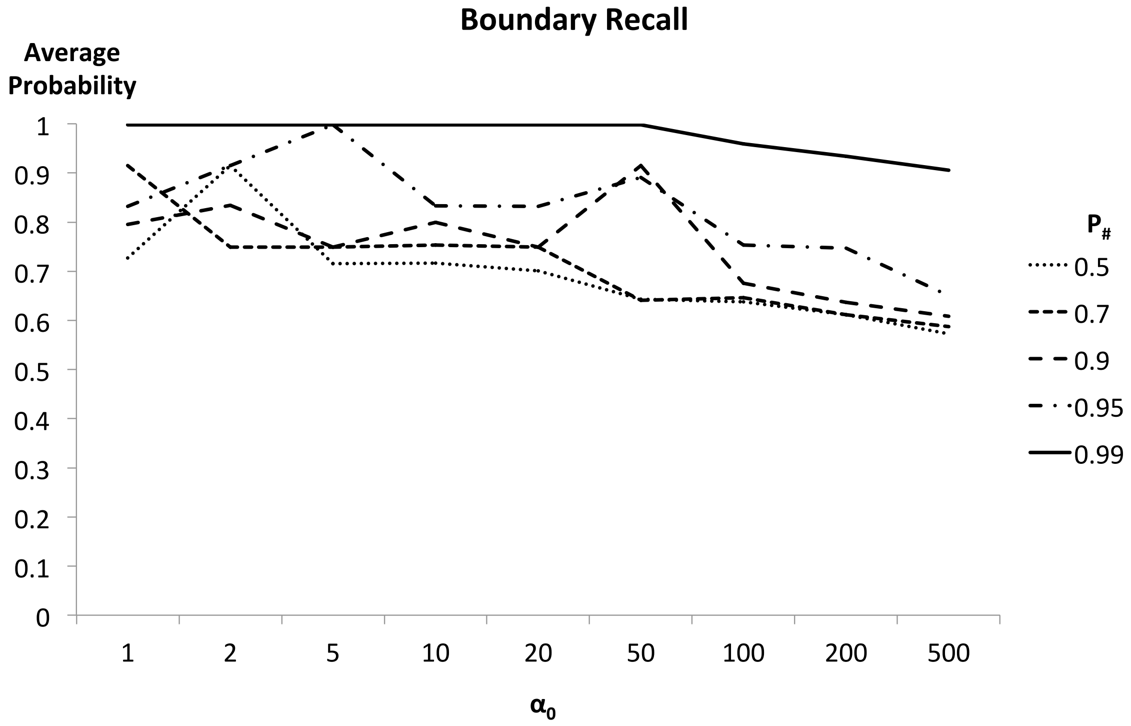


Figure 5. Average boundary recall probability for Experiment 1 model simulations. Results are shown for a range of values of the two segmentation parameters, $p_{\#}$ and α_0. The model has good recall performance across a broad range of parameter values, with perfect performance for all combinations of $p_{\#} = 0.99$ and $\alpha_0 \leq 50$

segmentation performance improves with smaller values of α_0 that favor a smaller vocabulary, and larger values of $p_{\#}$, favoring shorter actions. This makes sense, given that for these corpora the true vocabulary is four actions, each of which is three motions long. For $p_{\#} = 0.99$ and $\alpha_0 \leq 50$ the model consistently produces a perfect segmentation across samples, suggesting that the posterior distribution for these parameter values is highly peaked around the true segmentation.

As α_0 gets bigger and $p_{\#}$ gets smaller, the model's prior expectations shift towards a larger vocabulary, and longer actions. As a result, the model begins to *undersegment*, joining together adjacent actions. For instance if the true segmentation is `BLR TFD PEC` the model might produce `BLRTFD PEC`, treating `BLRTFD` as a single action. This can be seen as analogous to the part-words and part-actions in the statistical segmentation work described earlier – extracting actions that cross a boundary in the true segmentation.

Across a broad range of parameters, our model produces a very good segmentation of corpora similar to those used by Baldwin et al. (2008), and modeled after classic statistical word segmentation experiments, and a range of parameter values consistently produce a perfectly segmented sequence. These results confirm that the sequential probabilities available in the corpus can, in principle, be used for segmentation.

Experiment 1a: Online Segmentation of Statistical Actions

In this experiment, we had participants directly segment a corpus of statistically-determined actions in order to see whether, like our model, people can segment a sequence of actions using only statistical cues from the co-occurrences of motions within the sequence. As noted earlier, while previous work has demonstrated that people are sensitive to statistical cues in an action stream, this work had people provide judgments of discrete, already segmented sequences. Here we examine whether people can also use statistical cues to make online boundary judgments, in order to establish that these two methodologies for evaluating segmentation performance align.

Having established this correspondence in Experiments 1a and 1b, our remaining experiments will use discrimination measures to evaluate participants' segmentation judgments and causal inferences. This approach allows us to query participants about the appropriate hierarchical level of unit, and allows us to see whether they have extracted the statistical structure of the stimuli without making them aware of the segmentation task. This approach also has the advantage of allowing us to use exactly the same measure for judgments of familiarity, causality and coherence, and to evaluate participants' responses in cases where a single correct segmentation of the stimuli may not be possible – such as cases where only causal information is available, or cases of conflicting statistical and causal cues. Discrimination measures also provide us with information about the kinds of items people have extracted from an unfolding stream.

Method

Participants. Participants were 43 U.C. Berkeley undergraduate students who received course credit for participating. Participants were randomly assigned to view one of four exposure corpora. For each exposure corpus, three test corpora were created, making for 12 possible combinations, counterbalanced across participants.

Stimuli. Similar to Baldwin et al. (2008), we used 12 individual video clips of object-directed motions (referred to as *small motion elements* or SMEs in the previous work), to create four *actions* composed of three SMEs each (see Figure 4). The SMEs in this experiment are identical to those in Meyer and Baldwin (2011). As in previous work, SMEs were sped up slightly and transitions were smoothed using iMovie HD to make the exposure corpus appear more continuous.

We created a 23 minute exposure corpus by randomly choosing actions to add to the sequence, with the condition that no action follow itself, and that all actions and transitions between actions appear an equal number of times, resulting in 90 appearances of each action and 30 appearances of each transition. To ensure that none of our randomly assembled actions were inherently more causal or meaningful, we also created a set of four non-actions, each composed of three SMEs that never appear together in the first corpus, and assembled a second corpus using these sequences as the actions. Finally, we created two additional corpora, using the same actions but presented in a different order, to ensure that good segmentation performance was not limited to a particular order of presentation.

To create our test corpora, we excerpted three 10-minute (154 or 155 action) sections from each of the four exposure corpora. After viewing one exposure corpus, each participant was tested on an excerpt from the other corpus assembled from the same actions, so that they segmented a sequence they had not previously viewed. We limited participants to 10 minutes of segmentation due to the overall length of the experiment.

Procedure. Before the exposure corpus, participants were instructed only to pay close attention, and told only that they would be asked about the video after it had played.

After the exposure corpus, participants were directed to watch the test corpus and, while watching, to divide the video up into action units by pressing a key whenever they believed that a natural and meaningful unit of action had ended and new one was beginning (these instructions are modeled on those used by Zacks (2004)). The exact instructions were as follows:

> In the next part of the experiment you will be asked to divide an action sequence into natural and meaningful units of activity. What we mean by natural and meaningful is as follows. Imagine I remove a pen cap and then start writing with the pen. Those two actions go together in a sensible way. Now what if, instead, I remove the pen cap and then tie my shoe. I could do those actions together, but they don't really go together. We would like you to press a button whenever you think a sequence that goes together has ended and a new sequence is beginning.

As noted earlier, previous work has shown that people can segment at multiple hierarchical levels. In order to guide participants towards segmenting at the level of actions rather than SMEs. participants were also told:

> People can also vary the level at which they break up actions. For example, you could separate "uncap pen" from "begin writing" or you could lump them together into one larger action: "writing with pen". We would like you to press a button to mark the end of the largest units that seem natural and meaningful to you, those that are like "writing with pen".

Following the exposure corpus, the test video was preceded by a supervised trial with a practice video (consisting of entirely novel stimuli) that allowed participants to learn the interface, and during which the experimenter would clarify that participants did not need to mark a division after every SME, but instead should be marking meaningful actions one or more SMEs in length.

The experiment interface was a custom web application using HTML5, JavaScript, PHP, and MySQL. However, all participants participated from in-lab computers.

Results

Figure 6 shows the distribution of participants' raw key presses across actions. In order to interpret participant responses, each key press was subsequently assigned to a boundary between two SMEs. Preliminary plotting of the timing of key presses within SMEs revealed that presses were distributed bimodally (Figure 6), with one mode occurring at the boundary between SMEs, and the other occurring in the middle of the SME. Since participants were instructed to mark the boundary at end of each action, we assigned the presses in the middle of the motion to the preceding boundary.

Accordingly, key presses during the first three quarters of an SME were rounded to the preceding boundary, and only those in the last quarter of an SME were assigned to the upcoming boundary. Redundant key presses – presses beyond the first assigned to a particular boundary by each subject–were excluded from analysis. Finally, the boundary after the final SME and any key presses assigned to it were excluded from analysis. This was done because the test corpus video ended immediately after the final SME and, therefore, participants did not have a complete opportunity to mark a boundary after the final SME.

Participants marked a total of 5830 boundaries between units of action (excluding 198 boundaries by the criteria described above). Boundaries marked between the final SME of one action and the first SME of the next were considered correct and all other marked boundaries were considered incorrect (giving a chance level of $\frac{1}{3}$ if boundaries were distributed randomly). The precision (correct marked boundaries / total marked boundaries) of the group was 48.3%, performance significantly greater than would be expected by chance (two-tailed binomial test, $p < 0.0001$).

Individual participants marked an average of 136.7 boundary points each (min = 31

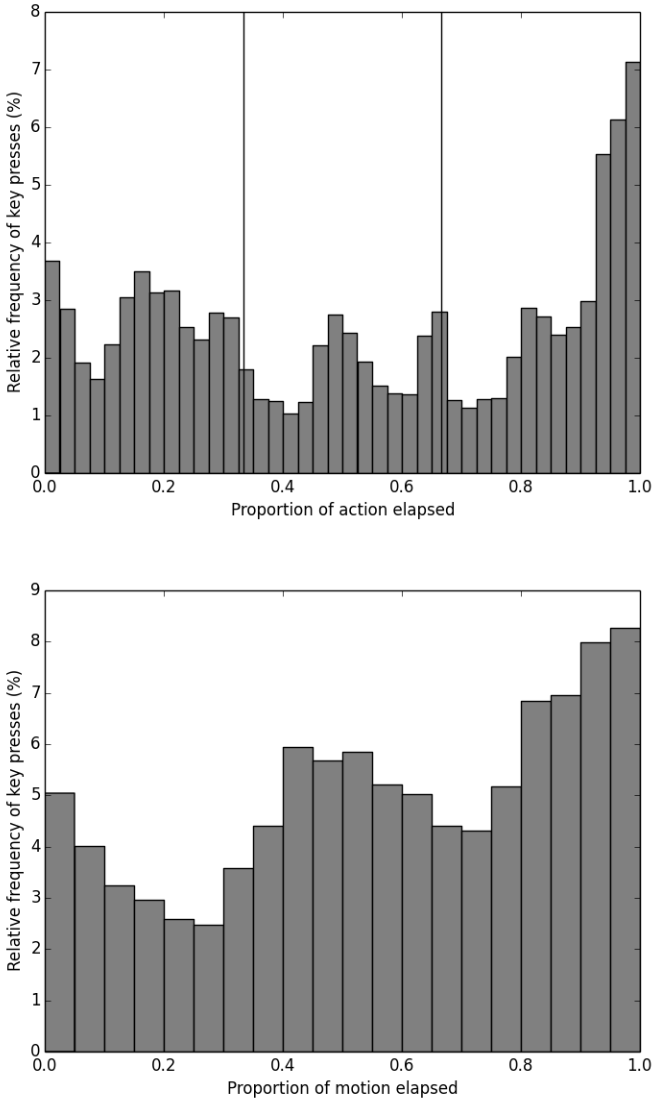


Figure 6. Top: Distribution of participants' boundary presses within actions. The two vertical lines represent the motion boundaries.The y-axis is the percentage of all presses occurring at that temporal point in their action. Bottom: Distribution of boundary presses within a motion. The y-axis is the percentage of all presses occurring at that temporal point in their motion.

max=374), out of approximately 465 possible boundaries and approximately 155 possible correct boundaries. The average precision among participants was 50.2%. 21 of the 43 participants performed significantly better than chance (two-tailed binomial tests, $p < .05$). The average recall (correct marked boundaries / number of correct boundaries in the test corpus) among participants was 42.7%. There was no effect of the particular actions viewed on participants performing above chance (Fisher's exact test, $p = 0.23$).

Discussion

The results of Experiment 1a are the first to demonstrate that people are able to provide online segmentation judgments for artificially constructed statistical actions, using the same type of segmentation paradigm previously applied to videos of everyday intentional actions (e.g., Newtson, 1973; Zacks, Tversky, & Iyer, 2001). Participants were significantly more likely to mark boundaries at the end of an action than within an action, indicating that they perceived actions, rather than parts of actions, to be the natural units in these sequences. Participants were able to provide correct segmentation judgments for these sequences using only the transitional probabilities between the motions in the sequence.

It is worth noting that participants in this task faced a significantly more challenging problem than our model: they were presented with continuous video to segment rather than a set of discrete motions, and they were not aware of the precise hierarchical level of segmentation we were looking for. For instance, segmenting each individual motion in the corpus (a valid level of segmentation) would lead to chance precision, while segmenting at a higher hierarchical level and grouping multiple actions would lead to low recall. Finally, since motions are only $\sim$1s long, participants needed to time their key presses very precisely to have them assigned to the correct boundary, whereas our model has no equivalent source of error.

Experiment 1b: Group Segmentation

The results of Experiment 1a suggest that people are capable of providing accurate online segmentation judgments for statistically constructed action sequences, and that this result is not dependent on the particular sequences of SMEs chosen to be actions, or the order in which they are presented in the corpus. Here, we present all participants with the same exposure corpus and test corpus, in order to use the aggregate boundary judgments to produce a "group segmentation", such as that used in past work on intentional action sequences (Newtson, 1973).

Method

Participants. Participants were 39 adults from the United States who were recruited through the Amazon Mechanical Turk marketplace (MTurk) and were paid $4.50 for participating. All participants were certified "Master Workers" on MTurk, indicating an especially high feedback score on their past work.

Stimuli. We selected one of the four exposure corpora from Experiment 1a and showed it to all participants. Similarly, all participants viewed the same test corpus (selected from the three test corpora from Experiment 1a matched to the selected exposure corpus). This gave us a significant number of responses to a single test corpus, allowing us to generate a meaningful "group-segmentation" of the test corpus.

Procedure. Online participants completed the same task as the in-lab participants in Experiment 1a, with a few small modifications. As online participants could not ask questions of – or receive feedback from–the experimenter, the instructions were expanded and clarified to address the most common questions and points of confusion observed in Experiment 1a. Specifically, participants were still asked to divide the video into "natural and meaningful units of activity" and were presented with the same examples. However, we replaced the instruction to "mark the end of the largest units that seem natural and meaningful" with the more explicit instruction "We are asking you to mark meaningful groups of (one or more) motions, we are **not** asking you to separate out each individual motion". To further minimize confusion, the video played during the interface practice session (between the exposure and test corpora) was changed to a short excerpt of the exposure corpus, instead of a clip of entirely novel stimuli.

Finally, whereas Experiment 1a had ended after the test corpus, there were a few attention-check questions after the test corpus in Experiment 1b. To assess how closely participants had been watching the exposure corpus, they were asked to report whether each of 16 motions (described in one sentence each) had appeared in the exposure corpus. Half of the described motions had appeared in the exposure corpus, and half had not. Experiment 1b used the same custom web interface as Experiment 1a.

Results

We used the same algorithm as in Experiment 1a to assign key presses to boundaries between SMEs, and excluded boundaries by the same criteria. Participants marked a total of 5961 boundaries between units of action (excluding 209 boundaries). 36 of 39 participants performed significantly better than chance (binomial test, $p < .05$) on the attention-check questions. Data from the three participants who did not perform significantly better than chance were excluded from analysis. The precision of the group was 41.2%, performance significantly greater than would be expected by chance (binomial test, $p < 0.0001$). The average precision among individual participants was 43.1%. 16 of the 36 participants performed significantly better than chance (binomial tests, $p < .05$). The average recall among participants was 42%.

Following Newtson (1973), we created a "group-segmentation" of the test corpus. For each boundary between SMEs, we counted the number of participants who marked a boundary at that location and calculated the mean, μ and standard deviation, σ, of these values. Then, we segmented the test corpus at each boundary at which the number of participants who marked a boundary was greater than $\mu + \sigma$. The precision of the group-segmented corpus was 77.4%, performance significantly better than would be expected by chance (two-tailed binomial test, $p < 0.0001$). The recall of the group-segmented corpus was 42.3%. We also created a group-segmented corpus based only on data from participants who individually performed significantly better than chance. This produced the group-segmented corpus seen in Figure C3. The precision and recall of this corpus are 91.4% (two-tailed binomial test, $p < 0.0001$) and 62.5%, respectively.

Discussion

These results further demonstrate that people's online boundary judgments are consistent with the statistical structure of the action sequence. The group segmented corpus from Experiment 1b shows quite good precision and recall, with the group segmentation produced by those participants who individually performed above chance (and therefore both understood and attended to the task, and segmented at a hierarchical level above that of the individual SME), approaching the levels of performance of our ideal observer model. Again, it is important to note that our ideal observer model provides a ceiling on possible performance, and that people performed well despite additional processing and response constraints not faced by our model. In addition, as can be seen in Figure C3, like our model results, the errors in this group segmentation are mostly undersegmentations – failing to divide two (or more) adjacent actions.

As noted earlier, previous work on statistical action segmentation has assumed that discrimination of actions versus part-actions and non-actions is the result of segmenting the exposure corpus appropriately, in order to extract coherent units. However, the correspondence between rating measures and boundary judgments had never been explicitly established. The results of this experiment verify that these methodologies do in fact align.

Statistical cues to causal structure

Just as people can recognize words from an artificial language, and distinguish them from non-words and part-words, we also know that they can recognize artificial actions grouped only by statistical relationships and can distinguish these sequences from non-actions (motions that never appeared together) and part-actions (motion sequences that cross an action boundary) (Baldwin et al., 2008). Here we investigate whether these groupings are also considered meaningful, and whether they are inferred to be causal.

Model Simulations

For this experiment, we used the same set of model simulations presented in Experiment 1. In addition to examining the model's segmentation directly, we can also query the model about what it has inferred about action sequences from this segmentation. In particular, here we examined the model's resulting action vocabulary – the set of sequences the model inferred as coherent units of behavior.

Figure 7 shows example results for the average probability that an action versus a part-action appears in the vocabulary. Notice that though the absolute difference varies with the parameter values, at least within the parameter ranges tested, actions are always more likely to be in the vocabulary than part-actions. In other words, across all parameter settings tested, after observing and segmenting the data, the model is more likely to believe that actions form coherent sequences that are likely to be observed again in the future, as compared to part-actions. Critically, this means that perfect segmentation performance is not required in order to extract and recognize statistically coherent units of action.

Finally, while there are no observed causal effects in these corpora, we can treat the effects as unobserved, and ask which sequences the model thinks would have been likely to lead to effects. In this case, the probability of a sequence leading to an effect reverts to the prior probability under the model, which is $(\pi \cdot \omega) + (1 - \pi)\epsilon$ for actions in the vocabulary and ϵ for other sequences. Therefore, when $(\pi \cdot \omega) + (1 - \pi)\epsilon > \epsilon$, the model will predict that actions are more likely to lead to effects than other motion sequences. This inequality will be true as long as $\omega > \epsilon$ – in other words, as long as effects are more likely to follow causes than non-causes. As $\omega >> \epsilon$, actions become increasingly more likely to be causal

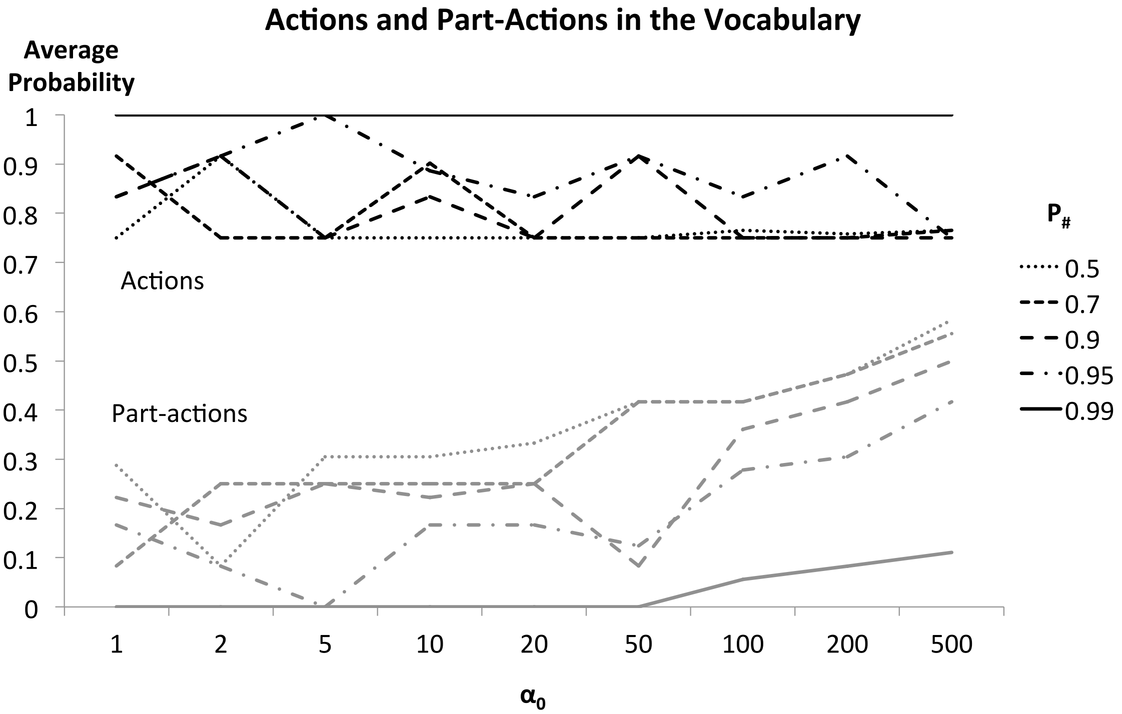


Figure 7. Average probability of a sequence appearing in the action vocabulary. Results are shown for a range of values of the two segmentation parameters, $p_{\#}$ and α_0. Across all parameter values, actions are more likely than part-actions

relative to part-actions and non-actions.

These results confirm that the sequential probabilities available in the corpus can in principle be used to distinguish actions from part-action and non-action sequences. The model also suggests that actions are more likely to be causal than other sequences, even when effects are not observed, and that the true action sequences used to construct the corpus are more likely to be identified as coherent units of motion that appear in the action vocabulary than sequences that cross an action boundary in the true segmentation, and that this inference about the relative coherence of actions vs other sequences can be made even without a perfect segmentation of the corpus. In the following two experiments, we examine people's inferences and segmentation performance for these same corpora.

Experiment 2: Coherence and Causality of Statistical Actions

We hypothesized that, like our model, participants would judge artificial *actions* to be more coherent and meaningful than similar *non-action* and *part-action* sequences (see Figure 4), and would also view *actions* as more likely to cause a (hidden) effect than *non-actions* and *part-actions*.

Method

Participants. Participants were 93 U.C Berkeley undergraduate students, who received course credit for participating. Participants were randomly assigned to view one of the two exposure corpora, and were also randomly assigned to one of three follow-up question conditions. 28 participants were assigned to the familiarity condition, 30 to the causal condition, and 37 participants were assigned to the coherence condition.

Stimuli. We used two of the exposure corpora used for Experiment 1. We also created four *action*, four *non-action* and four *part-action* comparison stimuli, where a non-action is a combination of three SMEs that never appear together in the exposure corpus, and a part-action is a combination of three SMEs that appears across a transition (e.g., the last two SMEs from the first action and the first SME from the second action, see Figure 4). Following Experiment 1, the actions of one corpus were used as the non-actions of the other corpus.

Procedure. All participants were instructed to attend closely to the exposure corpus, and were told that they would be asked questions about it later. Following the exposure corpus, participants in the *familiarity condition* were presented with all 12 actions, non-actions, and part-actions individually, and asked “How familiar is this action sequence?”. They responded by choosing a value on a 1 to 7 Likert scale, with 1 representing “not familiar” and 7 representing “very familiar” (other than the use of ratings instead of a forced choice format, this condition is almost identical to Baldwin et al., 2008). In the *causal condition*, participants were given a “hidden effect” cover story before viewing the exposure corpus. These participants were told that certain actions would cause the bottle being manipulated to play music, but that they would be watching the video with the sound off. Following the exposure corpus, these participants were asked "How likely is this sequence to make the bottle play a musical sound?", with 1 representing "not likely" and 7 representing "most likely". Finally, in the *coherence condition*, participants were asked the question "How well does this action sequence go together?". Like participants in Experiment 1, they were given the example of removing a pen cap and then writing with the pen as "going together" and of removing a pen cap and then tying your shoes as "not going together". As in Experiment 1a, prior to rating these sequences, participants in all conditions were first presented with two practice videos featuring sequences of motion that had not appeared in the exposure corpus, performed by a different actor, in order to become familiar with the ratings procedure. They then rated all test items on a scale with 1 being "does not go together" and 7 being "goes together well".

For all conditions, we used a custom Java program to present video of action sequences and collect ratings. The program presented all 12 actions, non-actions, and part-actions individually and in a random order.

Results

We analyzed results by condition, using 2×3 ANOVAs on exposure corpus (1 or 2) and sequence type (action, non-action, part-action). Results are shown in Figure 8.

In addition, in order to examine the effect of the particular sequences chosen as actions, we ran 2×8 ANOVAs on sequence type (action or non-action), and action (the 8 possible three-motion combinations used as actions and non-actions, see Figure 4). Part-actions were not included in this analysis since these sequences were unique to each corpus, and so did not change sequence type across corpora, and were not rated by all participants.

Ratings from 27 participants in the *familiarity condition* were analyzed (data from one additional participants who rated all sequences identically as either a 1 or 7 was discarded). As predicted by previous results (Baldwin et al., 2008; Meyer & Baldwin, 2011), there was an overall significant effect of sequence type $F(2, 50)= 25.14$, *MSE*= 41.12, $p < 0.0001$, with actions rated significantly more familiar than part-actions and non-actions $t(26)= 5.84$, $p < 0.0001$, one sample t-test on contrast values, and part-actions rated significantly more familiar than non-actions $t(26)= 3.65$, $p < 0.01$. There was no effect of exposure corpus $F(1, 25)= 0.16$, *MSE*= 0.42, $p = 0.69$, and no interaction between corpus and sequence type $F(2, 50)= 0.01$, *MSE*= 0.02, $p = 0.99$. There was also no effect of action sequence $F(7, 208)= 0.68$, *MSE*= 2.33, $p = 0.69$, and no interaction between action sequence and sequence type, $F(7, 208)= 0.37$, *MSE*= 1.26, $p = 0.92$.

Ratings from 29 participants in the *causal condition* were analyzed (data from one additional participant was discarded). As predicted, there was an overall significant effect of sequence type $F(2, 54)= 10.20$, *MSE*= 12.869, $p < 0.001$, with actions rated as significantly more likely to cause a musical effect than part-actions or non-actions $t(28)=$ 2.36, $p < 0.01$, one sample t-test on contrast values, and part-actions rated significantly more likely to be causal than non-actions, $t(28) = 2.36$, $p < 0.05$. There was no effect of exposure corpus, $F(1, 27)= 2.21$, *MSE*=4.6, $p = 0.15$, and no interaction between corpus and sequence type, $F(2, 54)= 0.56$, *MSE*= 0.70, $p = 0.57$. There was also no effect of action sequence, $F(7, 224)= 1.66$, *MSE*= 6.09, $p = 0.12$, and no interaction between action sequence and sequence type, $F(7, 224)= 1.51$, *MSE*= 5.52, $p = 0.17$.

Ratings from 37 participants in the *coherence condition* were analyzed. As predicted, there was an overall significant effect of sequence type $F(2, 70)= 9.18$, *MSE*= 14.47, $p < 0.001$, with actions rated as going together significantly better than part-actions or non-actions $t(36)= 3.87$, $p < 0.001$, one sample t-test on contrast values. There was also a marginally significant difference between part-action and non-action ratings $t(36)= 2.0$, $p = 0.05$. There was an effect of exposure corpus,$F(1, 35)= 7.3$, *MSE*= 14.47, $p < 0.05$ but no interaction between corpus and sequence type, $F(2, 70)= 0.44$, *MSE*= 0.70, $p = 0.65$.

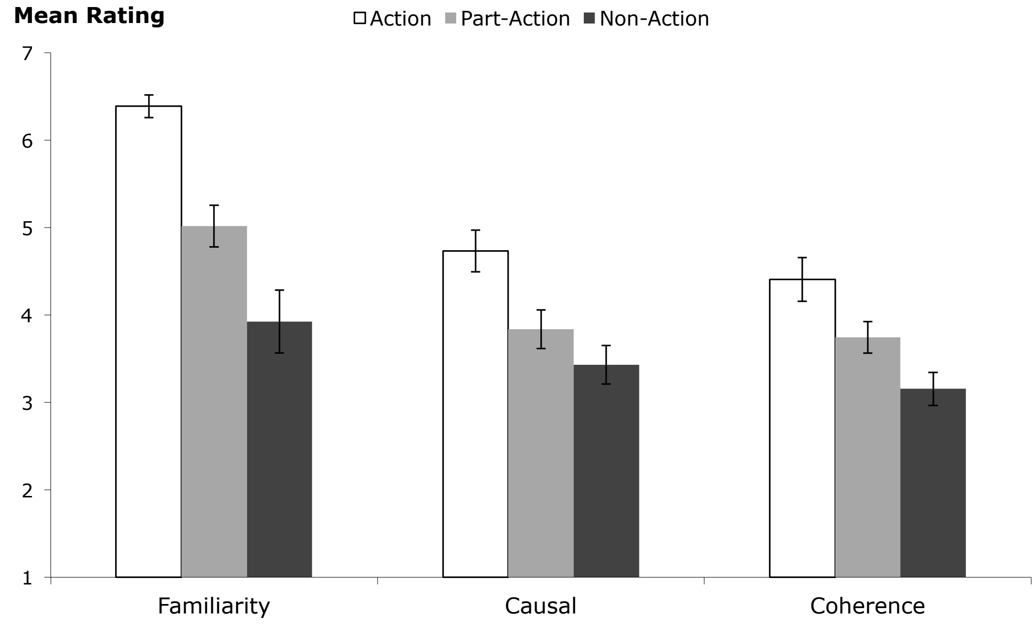


Figure 8. Results of Experiment 2. Error bars show one standard error.

Similarly, there was a marginal effect of action sequence $F(7, 280) = 1.99$, $MSE = 7.38$, $p = 0.057$, but no interaction between action sequence and sequence type, $F(7, 224) = 0.86$, $MSE = 3.18$, $p = 0.54$.

Discussion

The results of this experiment support the hypothesis that people experience sequences of action grouped only by their statistical regularities as causally significant, meaningful groupings. Participants rated actions as more likely to cause a hidden musical effect than part-action and non-action sequences. Similarly, participants rated actions as going together (a question we used as a measure of sequence coherence and meaningfulness) significantly better than other sequences. Anecdotally, a number of participants reported a feeling that the action sequences made more intuitive sense to them than the other sequences. Finally, these results support the findings of Experiment 1, and of Baldwin et al. (2008), that people are able to extract and recognize statistically grouped actions from within a longer action sequence, and differentiate them from other non-action groupings.

It is important to note that participants rated actions as more familiar, more likely to be causal and more coherent than part-action and non-action sequences, even though all sequences were equally arbitrary, and in fact the non-actions for one exposure corpus were the actions for the other, meaning that the same sequences reversed their rating merely based on the number of times the SMEs appeared together. While the fact that there was a marginal effect of sequence in the coherence condition suggests that some sequences were individually perceived as more or less meaningful, the perception of all sequences was nonetheless influenced by being an action or a non-action in the same way. Across conditions, the lack of interaction between the action sequences and their sequence type confirms that it was not the particular choice of motion combinations that led participants to rate actions as more familiar, causal and coherent. In fact, across conditions, all 8 individual sequences had a higher average rating when they were actions than when they were non-actions.

These results have several important implications. First, they demonstrate that people's sensitivity to the statistical patterns in the exposure corpus is not simply an artifact of the impoverished stimuli, but appears to play a real role in their subsequent understanding of the intentional structure of the action sequence. The fact that participants found the statistically grouped actions to be more coherent suggests that they do not experience the sequences they segment out as arbitrary, but assume that they are meaningful groupings that play some (possibly intentional) role. This is further supported by the results from the causal condition which show that, even without being presented with overt causal structure, people believe the statistically grouped actions are more likely to lead to external effects in the world.

Finally, these results also support our hypothesis that inference of action structure and causal structure are linked, with statistically grouped actions being perceived as more likely to also be causal variables. This result is consistent with our computational model, which also predicts that, without other evidence of causal structure, actions are more likely to be causal than non-action and part-action sequences.

Using Causal Structure

Our previous results suggest that statistical action structure can be used to infer causal relationships, but can causal relationships be used to identify meaningful actions?

To investigate this, we constructed two new exposure corpora. In these corpora, there were no *a priori*, statistically-grounded actions. Instead, each exposure corpus was assembled using four SMEs, so that each individual SME would be seen an equal number of times, and all possible length three sequences of SMEs would also occur with equal frequency.[2] Throughout the exposure corpus, no length three subsequences containing repeats of an SME were allowed to occur. This resulted in 24 possible three-motion sequences ("triplets"). A target triplet of SMEs was then randomly chosen as the "cause". Whenever this sequence of motions was performed in the exposure corpus, it was followed by an observable causal outcome (see Methods section of Experiment 3a, as well as Appendix B for further details on the construction of these corpora). We first ran the model on the exposure corpora to examine its segmentation and causal inference performance, and then looked at human performance on these same corpora.

Model Simulations

As before, an abstract representation of each unsegmented corpus was used as input to the model (see Appendix C for an example corpus), with a letter standing for each SME, and "*" representing a causal outcome. The model has three causal inference parameters:

[2] In fact, it turns out that all length two sequences appear with approximately equal frequency as well, so that the transitional probability between any two motions is ≈ 0.33

π the probability that an action is causal, ω the probability that a causal sequence leads to an effect, and ϵ the probability of an effect following a non-causal sequence. However, as noted earlier, what is relevant for causal inference is the ratio of ω to ϵ. Therefore, we use a fixed value of $\epsilon = 0.001$, and vary only ω. We also maintained the relationship $\omega >> \epsilon$, reflecting our assumption that causes are relatively effective, and are much more likely to precede effects than are non-causes. Following the results of Experiment 1, we used segmentation parameter values $\alpha_0 = 3$ and $p_{\#} = 0.99$ throughout,[3] and evaluated model results across values of $\omega \in \{0.5, 0.7, 0.9, 0.99\}$ and $\pi \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

We were interested in seeing whether the model could infer the correct causal subsequence from within the longer sequence of motions (since there was no “correct” segmentation for the remainder of the corpus, overall segmentation performance is not evaluated here). For all parameter values tested, the model performed perfectly or almost perfectly, correctly identifying the target triplet as causal across all samples and parameter values, $p(c_{target} = 1) = 1$. Additionally, for all parameter values, the model correctly segmented either 23/24 or 24/24 occurrences of the causal triplet. Similarly, the target triplet appeared in the vocabulary across all samples.

We ran our model on corpora designed so that all possible three-motion sequences occurred equally often. Across all parameter values, our model consistently identifies the correct causal sequence and segments out this sequence as an action. This suggests that, even when the transitional probabilities between motions are uniform, it is possible to identify and extract the correct length causal sequence, using only the causal statistics in the corpora.

Experiment 3a: Inferring causal variables

This experiment investigates whether people are able to pick out causal subsequences from within a longer stream of actions, and whether they use this causal information to

[3] Additional simulations confirmed that causal inference results were not significantly affected by changing the segmentation parameters.

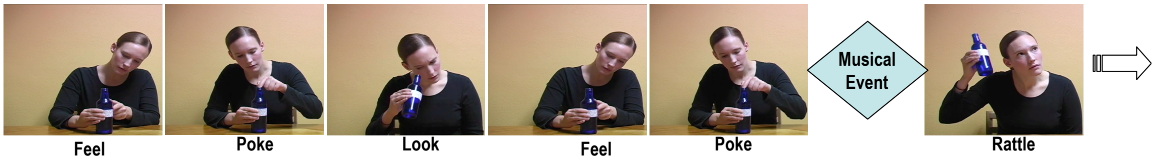


Figure 9. A portion of one of the Experiment 3 exposure corpora. four SMEs `Poke, Look, Feel, Rattle` are distributed so that all possible triplets appear equally often. A target triplet `Look, Feel, Poke` is chosen to cause a sound.

inform their action segmentation. Specifically, we hypothesized that when statistical cues to action segmentation are unavailable, adults, like our model, will be able to use causal event structure to identify meaningful units of action.

Method

Participants

Participants were 95 U.C. Berkeley undergraduates who received course credit for participating. 30 participants were assigned to the familiarity condition, 30 to the causal condition and 35 participants were assigned to the coherence condition.

Stimuli. The structure and stimuli for this experiment closely matched that of Experiment 2. However, in Experiment 3, the corpora were specially constructed so that all possible combinations of three motions appeared equally often together, so that joint and transitional probabilities could not be used to identify groupings (see Figure 9). Throughout the exposure corpus, no length three subsequences containing repeats of an SME were allowed to occur. This resulted in 24 possible SME triplets. A target triplet of SMEs was then randomly chosen as the "cause". Whenever this sequence of motions was performed in the exposure corpus, it was followed by the object playing music (participants were able to hear the music, unlike in Experiment 2).

The exposure corpus was created by first generating 24 shorter video clips. Each clip was designed to have a uniform distribution of both individual SMEs and of SME triplets. Specifically, in each clip, the four unique SMEs appear exactly six times each, and 23 of the

24 possible SME triplets appear exactly once each. We designed the 24 clips by using a De Bruijn sequence (van Aardenne-Ehrenfest & de Bruijn, 1951), a cyclical sequence within which each subsequence of length n appears exactly once as a consecutive sequence (see Appendix B for algorithmic details). These 24 video clips were shown consecutively in the exposure corpus, but were clearly separated from each other by text notifying the participant of the beginning and end of each shorter clip. The result was an exposure corpus composed of 24 short video clips, with each SME appearing 144 times throughout the complete corpus, and each triplet appearing 20 to 24 times.

iMovie HD was used to assemble the exposure corpus and add a cartoon sound effect following every appearance of the target sequence. Two different exposure corpora, each using a distinct set of four SMEs were created. `Look, Poke, Feel`, and `Rattle` were used to create the first exposure corpus, with `Look Feel Poke` being the target triplet, and `Read, Slide, Blow`, and `Empty` were used to create the second exposure corpus, with `Slide Blow Empty` being the target triplet.

Procedure. Participants were divided into the same three rating conditions as in Experiment 2. The procedure was identical to that of Experiment 2, with the difference that after viewing the exposure corpus, participants rated all 24 possible SME triplets, and that all participants were told that certain action sequences caused the bottle to play music.

Results

We analyzed all results using 2×2 ANOVAs on exposure corpus (1 or 2) and sequence type (target, other). No effects of exposure corpus were found. Results are shown in Figure 10.

Ratings from 28 participants in the *familiarity condition* were analyzed (data from an additional two participants who rated all sequences identically as either a 1 or 7 was discarded). There was no effect of sequence type $F(1, 26) = 1.58$, $MSE = 1.74$, $p > 0.22$. Participants rated the target sequence and the other SME triplets as equally familiar.

Ratings from 30 participants in the *causal condition* were analyzed. As predicted, there was a significant effect of sequence type $F(1, 28)= 193.97$, $MSE= 310.439$, $p < 0.0001$, with the target sequence being rated as much more likely to lead to a musical sound than the other SME triplets. This difference remains significant when the target sequence is compared only to non-target permutations of the same motions, $t(29)= 10$, $p < 0.0001$, one sample t-test on contrast values.

Ratings from 35 participants in the *coherence condition* were analyzed. As predicted, there was a significant effect of sequence type $F(1, 33)= 19.44$, $MSE= 47.1$, $p < 0.0001$, with the target sequence rated as going together significantly better than the other SME triplets. This difference remains significant when the target sequence is compared only to non-target permutations of the same motions, $t(34)= 4.18$, $p < 0.001$, one sample t-test on contrast values.

Finally, we compared ratings from the causal and coherence conditions using a 2×2 ANOVA on condition and sequence type. There was no effect of condition $F(1, 63)= 1.04$, $MSE= 1.94$, $p = 0.31$, but a significant interaction between condition and sequence type $F(1, 63)= 32.1$, $MSE= 68.37$, $p < 0.0001$. Post-hoc t-tests found that causal ratings were significantly greater than coherence ratings for the target sequence, $t(63)= 2.98$, $p < 0.01$, two sample t-test, while coherence ratings were significantly greater than causal ratings for the non-target sequences $t(63)= 5.95$, $p < 0.0001$.

Discussion

This experiment is one of the first to demonstrate that people can infer a correctly ordered set of causal variables from within a longer temporal sequence. In fact, the results of this experiment suggest that it was a relatively easy task for participants. Participants in the causal condition were nearly at ceiling in their ratings of how likely sequences were to lead to a musical effect, with the target sequence having a mean rating only slightly below 7 and the remaining sequences being rated a bit below 2. Importantly, this was true even when the non-target sequences contained all the correct motions, but in an incorrect order.

The results of this experiment also provide further support for a relationship between action segmentation and causal inference. Even though there were no statistically grouped actions in this experiment, participants still perceived the target sequence as being more meaningful (going together better) than the other sequences, suggesting they had nonetheless segmented it out as a coherent action unit. It is worth noting that the ratings for the coherence question were different than those for the causal question, suggesting that participants did interpret the question as one of meaningfulness, rather than an alternate phrasing of the causality question. Note that this difference is qualitatively predicted by the model – an artifact of not only identifying the causal triplet but of segmenting the corpus, is that other triplets end up in the action vocabulary as well, but without the consistency of the target triplet. In contrast, other triplets are *never* identified as causal.

Finally, while not directly relevant to our model predictions, it is interesting to note that, despite correctly identifying the target sequence as causal, participants *did not* rate it as more familiar than the other sequences. Instead, participants appeared to be aware that they had seen all the sequences an equal number of times, and rated them all as equally familiar. This implies that participants are not judging the target sequence as more coherent or more likely to be causal due to some sort of low level saliency effect that causes them to remember this particular sequence more clearly. It also suggests that participants, at least in this context, interpret the familiarity question as a question about frequency of appearance. Previously, Meyer and Baldwin (2011) found that familiarity responses reflect both a sequence's overall frequency of appearance, and the conditional probabilities within the sequence. In this case, both conditional probabilities and overall frequency were balanced across all SME triplets, and neither provided a cue to which sequences were more meaningful. These results suggest that participants may be aware that certain sequences are more causal or more coherent, while also being aware that they have seen other sequences equally often, and that coherence may more closely correspond to the units our model extracts than familiarity. Therefore, this pattern of results is compatible with our model predictions if we treat "how familiar is this sequence" as a question about the frequency of that sequence's appearance (or perhaps about the conditional probabilities within the sequence), and "how coherent or meaningful is this sequence" as a more holistic question about the extraction of that sequence as a meaningful unit, independent of its statistical probabilities.

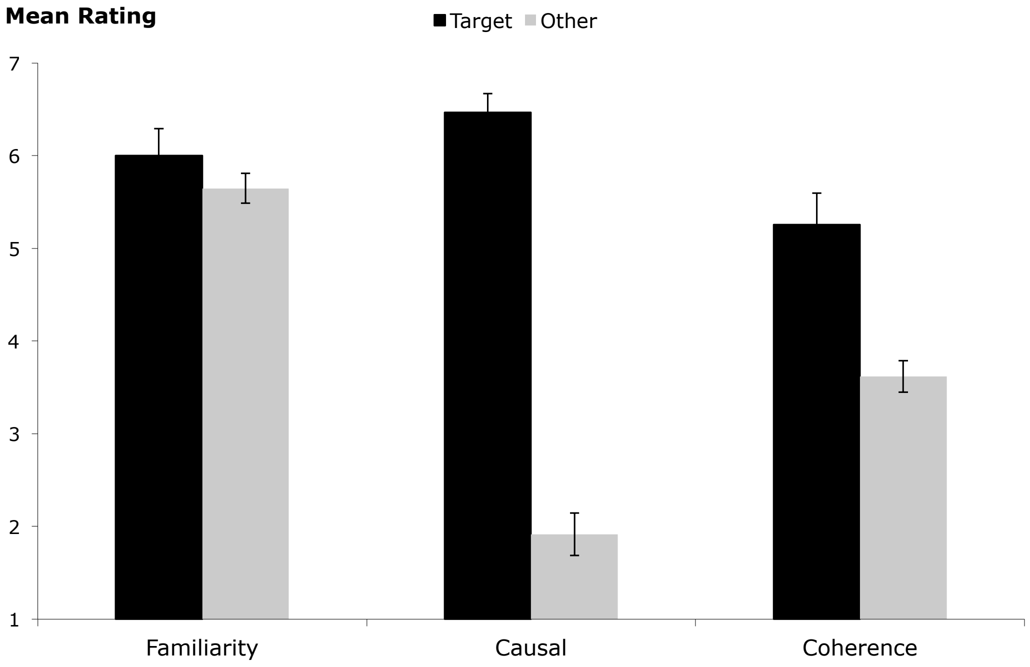


Figure 10. Results of Experiment 3a. Error bars show one standard error.

Experiment 3b: Identifying Correct Causal Subsequences

Experiment 3a found that people are able to pick out causal subsequences from within a longer stream of actions. However, in the previous experiment, we only asked participants to rate actions composed of three SMEs each. This potentially leaves open the question of whether, like our model, participants are really identifying the correct causal subsequence, or whether they might actually prefer a subsequence or supersequence of the causal actions if given the choice. In this experiment, we explicitly look at whether participants are able to identify causal subsequences of the correct length.

Method

Participants

Participants were 53 U.C. Berkeley undergraduates who received course credit for participating.

Stimuli. The structure and stimuli for this experiment closely matched those of Experiment 3a. The same exposure corpora were used as in Experiment 3a. As in Experiment 3a, `Look Feel Poke` was the target triplet in the first exposure corpus, and `Slide Blow Empty` was the target triplet in the second corpus. In both corpora, the target triplet was always followed by a cartoon sound effect.

A series of 30 test stimuli were created for each exposure corpus. Each test stimulus was an action composed of between 1 and 5 SMEs. For each corpus, the set of test stimuli was constructed so that it contained the target action, single and double length subsequences of the target action (e.g., for `Slide Blow Empty` these would be `Empty` and `Blow Empty`), and quadruplet and quintuplet length supersequences of the target action (e.g., `Read Slide Blow Empty` and `Empty Read Slide Blow Empty`). There were also non-target actions of each of these lengths, some of which contained subsequences of the target action (e.g., `Read Blow Empty`) and others that did not (e.g., `Empty Slide Read`). See Figure 11 for a complete set of test sequences for one corpus. iMovie HD was used to assemble the test stimuli, as in Experiments 1 and 2.

Procedure. Participants were randomly assigned to view one of the two exposure corpora. All participants were given viewing instructions identical to those in the causal condition of Experiment 3a. After viewing the exposure corpus, participants were asked "If

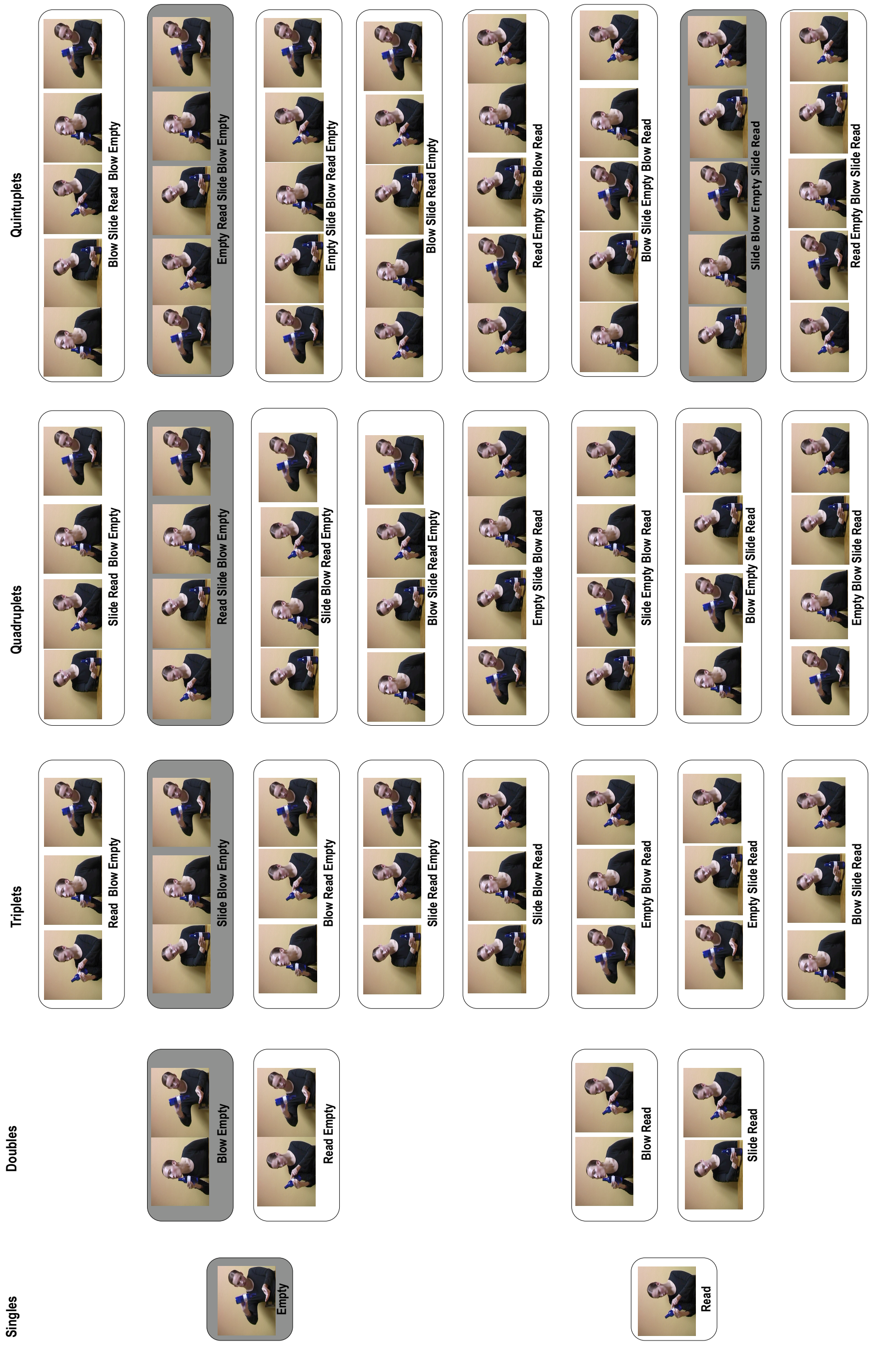


Figure 11. The 30 test stimuli for Experiment 3b, exposure corpus 1. Test stimuli include both subsequences and supersequences of the target action `Slide Blow Empty`. These are shown with a dark background. Comparison sequences of equal length are shown with a light background.

you were trying to cause a musical sound, how likely would you be to use this action sequence?”, for all 30 test sequences. As in the causal condition of Experiments 2 and 3a, participants responded by choosing a value on a 1 to 7 Likert scale, with 1 representing “not likely” and 7 representing “very likely”.

Results

We analyzed results using 2×2 ANOVAs on exposure corpus (1 or 2) and sequence type (contains target, other). No effects of exposure corpus were found. Results are shown in Figure 12.

Ratings from all 53 participants were analyzed. There was a significant effect of sequence type $F(1, 51)= 248.61$, $MSE= 1714.12$, $p < 0.0001$, with sequences containing the target being rated as much more likely to lead to a musical sound than other sequences.

We compared ratings of each sequence containing the target to the non-target sequences of equivalent length, using one sample t-tests on contrast values. In all cases, the target-containing sequences were rated as significantly more causal than the non-target sequences of the same length. Triplet sequences: $t(52)= 13.60$, $p < 0.0001$. Quadruplet sequences: $t(52)= 14.46$, $p < 0.0001$. Quintuplet sequences: $t(52)= 12.85$, $p < 0.0001$.

We compared ratings of single and double length terminal subsequences of the target to other single and double length sequences, using one sample t-tests on contrast values. Subsequences of the target were rated as significantly more causal than other sequences of equivalent length. Single SME sequences: $t(52)= 3.64$, $p < 0.001$. Double sequences: $t(52)= 5.96$, $p < 0.0001$.

Finally, we compared ratings of the target triplet to subsequences and supersequences of the target, using one sample t-tests on contrast values. Subsequences were rated as significantly less causal that the target triplet itself. Single SME subsequence of target: $t(52)= 15.08$, $p < 0.0001$. Double subsequence of target: $t(52)= 14.80$, $p < 0.0001$. Finally, the target sequence and supersequences of the target were rated as equally causal.

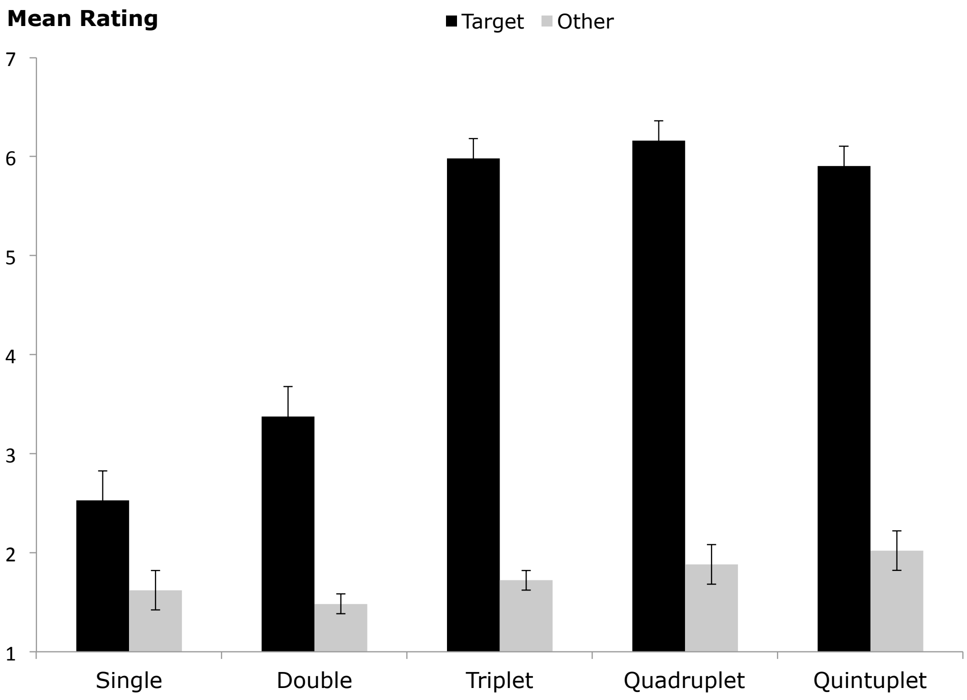


Figure 12. Results of Experiment 3b. Error bars show one standard error.

Quadruplet sequences: $t(52)= 1.19$, $p > 0.24$. Quintuplet sequences: $t(52)= -0.32$, $p > 0.74$.

Discussion

The results of Experiment 3b confirm that people can identify the correct length causal sequence from within a longer sequence. Participants correctly rated only sequences containing the complete target triplet as very likely to lead to the effect. They were able to distinguish these causal sequences from sequences containing only a subset of the necessary sequence, and from sequences containing all of the same motions, but in an incorrect order.

Once again, there are a number of intriguing results that, while not directly relevant to our model predictions, are worth noting. First, participants rated subsequences of the causal triplet as less likely to be causal than the full triplet, but as more likely to be causal than other sequences of equivalent length perhaps suggesting that participants recognized these subsequences as necessary but not sufficient for producing the effect. Second, while the jump in causal ratings for the target triplet and its supersequences confirms that, like our model, participants recognized the triplet sequence length as both necessary and sufficient for causing the effect, it is interesting to note that participants did not penalize the supersequences for including superfluous motions. Both results may reflect a bit of ambiguity in the question we asked participants. While our model was tasked with inferring only the exact causal sequence, participants may have wanted to also signal their knowledge that the subsequences were necessary (though not sufficient), and that the supersequences, while including superfluous actions, would nonetheless lead to the effect. Both of these results are compatible with our model results. In fact, interpreting the question as asking how likely this sequence would be to be followed by the effect, would lead to the same pattern of results seen in the human data. It is important to emphasize that this pattern of results would not be expected to occur if participants thought that a supersequence was the causal sequence, with the causal triplet being necessary, but not sufficient. In that case, the target triplet would be expected to rate higher than non-target triplets, but significantly lower than its causal supersequence.

Evaluating Alternative Models

Experiments 1, 2, and 3 together suggest that people can use statistical structure in action sequences to inform their causal inferences, and can use causal relationships between actions and external outcomes to help determine action structure. In particular, they demonstrate that people believe that action sequences with higher internal transition probabilities are good candidate causes, and that sequences of motions that predict outcomes in the world are likely to group together, independent of their transition probabilities.

However, before we can attribute this performance to joint causal and statistical inference, we must look at whether other simpler heuristics could potentially explain these results as well. Below we review a number of alternative approaches to segmentation and causal inference and explore which aspects of the previous experiments they account for. We next present a new experiment to differentiate these alternative models from our joint inference model.

Alternative Models

Transitional Probability Model. One alternative segmentation model that has been looked at in the statistical word learning literature (e.g., Saffran, Newport, & Aslin, 1996) is a model that operates directly on the conditional probabilities between syllables, without an explicit generative model of words or sentences (or in our case, of actions or action sequences). Following Frank, Goldwater, Griffiths, and Tenenbaum (2010) we can represent such a model by computing the conditional probability of each motion given the preceding motion from occurrence counts in the corpus

$$p(m_i \mid m_{i-1}) = \frac{C(m_{i-1}, m_i)}{C(m_{i-1})} \tag{7}$$

where $C(m_{i-1}, m_i)$ is the number of times the motions have occurred in sequence, and $C(m_{i-1})$ is the number of times m_{i-1} has occurred overall. We can compute this value for all possible pairs of SMEs in the corpus, and then insert a boundary in the corpus whenever a local minima in transitional probability occurs.

In the Experiment 1 and 2 corpora, conditional probabilities are 1 between motions in an action, and 0.33 between actions. For example, for the sequence `PEC USA` $p(\mathrm{C} \mid \mathrm{E}) = 1$ while $p(\mathrm{U} \mid \mathrm{C}) \approx 0.33$. By inserting a boundary any time $p(m_i \mid m_{i-1}) < 1$, this approach would correctly segment the corpus into its component actions. While this model has no causal component, it is perhaps plausible to think that, in the absence of causal data, people could first use transitional probabilities to segment the sequence, and then identify the actions they segmented as potential causes afterwards.

However, this conditional probability heuristic would fail in Experiment 3, since in these corpora the conditional probability between all pairs of SMEs is identical, ≈ 0.33, and so conditional probability would be unable to identify the target sequence. This remains true even if we treat the musical event as just another statistical occurrence (another "motion") in the sequence, since the motion preceding the effect is not sufficiently predictive on its own, and has a low probability of transitioning to the effect (≈ 0.11). Therefore, transitional probabilities between motions cannot account for the results of Experiment 3.

Causal Inference Only Model. Another possibility for accounting for the Experiment 3 data is a purely causal inference model, that does not take the statistical co-occurrences of the actions into account. As mentioned in the discussion of Experiment 3, people could evaluate the causal strength of various sequences by using the conditional probability of the effect following that sequence in the corpus.

One disadvantage to this option is that it does not answer the question of how to pick out which sequences to evaluate (the causal variable problem). One plausible heuristic would be to only consider sequences up to a certain length, for instance no more than 5 motions long. In the case of 4 possible motions (as in Experiment 3), this gives us 1364 potential causal sequences to consider. We can then compute

$$p(\text{event} \mid \text{sequence}) = \frac{C(\text{sequence}, \text{event})}{C(\text{sequence})} \tag{8}$$

For each of these sequences. this approach would give us $p(\text{event} \mid \text{target single}) \approx 0.16$, $p(\text{event} \mid \text{target double}) \approx 0.5$,
$p(\text{event} \mid \text{target triplet}) = p(\text{event} \mid \text{target quadruplet}) = p(\text{event} \mid \text{target quintuplet}) = 1$, and 0 for all other sequences, a result qualitatively very similar to human performance.

Since there is no segmentation component to this model, it makes no prediction for the Experiment 1 and 2 corpora, and cannot account for the human data on its own. However, it is possible that, in the presence of statistical patterns in the motion and no causal data, people apply a statistical learning strategy, such as the transitional probability strategy described above, while in the presence of causal data and no statistical action structure, they apply a simple causal learning strategy such as this one, without using joint inference in either case.

Sequential Structure Only Model. Another possibility is that our generative model of sequential structure is sufficient on its own, without needing the additional causal component. We already know that this is the case for segmenting the Experiment 1 and 2 corpora, where no causal data is available. For Experiment 3, we can once again treat events as simply another "motion" in the sequence. In this case, the model performs as well as our joint inference model, segmenting out all occurrences of the causal triplet. Once again, if people use a similar approach, they could first use this purely statistical approach to identify the target sequence, and only subsequently infer that it is more likely to be the cause.

Joint Inferences from Conflicting Cues

In order to discover whether people in fact perform joint inference, we need a case where our joint inference model and the alternative sequential inference models described above make distinct predictions.

We can test whether causal structure and action structure are jointly inferred by generating a new set of corpora, in which statistical and causal cues are both present, and are in conflict. As in the Experiment 1 corpora, there are four statistically determined actions, but in these new corpora there is also a target part-action that consistently leads to the object playing a musical sound.

The corpora used in this next set of simulations were exactly the same as those used in Experiment 1, but with an effect added after every occurrence of the target part-action (see Appendix C).[4] As before, an abstract representation of each unsegmented corpus was

[4] Note that this means that if, for example, the actions are `TFD` and `BLR`, and an effect is added so that we see `TFDBL*R` every time these two actions occur together, then treating the causal sequence as any of `TFDBL`,

used as input to the model, with a letter standing for each SME.

We first look at the predictions generated by our alternative models that perform causal inference and action segmentation separately, and compare these to our model's joint segmentation and causal inference results, to see whether joint inference produced distinct predictions, and whether those results were an improvement relative to those generated by the separate inference processes. Finally, we looked at human performance on these same corpora.

Conditional Probability. As described earlier, we can segment this corpus using the conditional probability of each motion given the preceding motion, inserting a boundary when this probability drops below some threshold. If we simply ignore events, then this approach would correctly segment the corpus into its component actions (as in Experiment 1), but would never extract the target part-action as a coherent unit.

We can instead treat the event as a motion. For instance, if the causal part-action is `TFDBL*`, composed of actions `TFD` and `BLR`, then $p(* \mid L) = 0.33$, $p(R \mid L) = 0.66$ and $p(R \mid *) = 0.66$, with the other conditional probabilities staying the same as before. If we set our segmentation threshold to $p(m_i \mid m_{i-1}) < 1$ as before, we get the segmentation `TFD BL *R` whenever the part-action occurs and `BL R` whenever the action `BLR` occurs outside this context. This approach never identifies the target part-action, and identifies only 3 out of 4 actions.

Alternatively, if we set the threshold to $p(m_i \mid m_{i-1}) < 0.66$, we will again recognize all four actions, but still get the segmentation `TFD BL *R` whenever the part-action occurs, and never identify the target part-action. Therefore, the conditional probability model would be unable to identify the target part-action as a coherent unit, or to propose it as a potential causal variable for sequential inference.

Causal Inference Only Model. We can again consider the conditional probability of motion sequences predicting the effect, for all possible sequences of 5 or fewer

`FDBL` and `DBL` is equally valid for this corpus, since they will all always co-occur with the effect.

motions. This approach would correctly identify the target part-action as causal $p(\text{event} \mid \text{target}) = 1$. It would identify all other actions, part-actions, and non-actions as non-causal, $p(\text{event} \mid \text{other}) = 0$, and has no mechanism of distinguishing actions from other non-causal sequences, treating them identically.

Sequential Structure Only Model. We can also look at the results of running our model using only the statistical structure of the motion sequence, and ignoring the causal information. This is equivalent to running the model on the Experiment 1 corpora, where there was no causal information present. As before, this model would correctly identify actions, and, depending on the parameter settings, would also discover part-actions. However, this model makes no distinction between the target part-action and other part-actions, and across all parameter values, never segments out the exact causal sequence.

We can again look at what happens in this model if we treat effects as another event in the sequence. In this case, the model will always come up with segmentations that embed the event. For instance, if the causal part-action is `TFDBL*`, composed of actions `TFD` and `BLR`, this model will produce either `TFDBL*R` or `TFD BL*R` depending on the parameter settings, but will never produce e.g., `TFDBL* R` or `TF DBL* R`. Therefore, this model has no mechanism for distinguishing the target part-action from other part-actions.

Joint Inference Results. We ran our model for the parameter range $\alpha_0 \in \{1, 2, 5, 10, 20, 50\}$ and $p_\# = 0.99$, the parameter range producing perfect segmentation performance in Experiment 1. Preliminary analysis indicated that, as in Experiment 1, results were not significantly influenced by the particular choice of α_0 within this range, so we present result collapsed across values of α_0.[5].

Since the exact values of the causal parameters π and ω had no impact in Experiment 2 within the range tested, we used a slightly coarser grid to reduce computation time: $\omega \in \{0.5, 0.7, 0.9, 0.99\}$ and $\pi \in \{0.01, 0.05, 0.1, 0.3, 0.5\}$.

[5] Collapsing in this way gives us $\sum_{\alpha_0} p(h \mid d, \alpha_0) p(\alpha_0)$, so that we are still sampling from the posterior distribution, with a uniform prior $p(\alpha_0)$ over the sampled values of α_0

Overall, the model succeeded at identifying the true actions, which appeared in the vocabulary with a probability ranging from 0.95 to 1 across parameter values. Similarly, the model consistently identified the target sequence,[6] which appeared in the vocabulary with average probability ranging from 0.5 to 1 across parameters. Whenever the target sequence appears in the action vocabulary, the model also correctly identifies it as causal. In contrast, sequences containing other part-actions appeared in the vocabulary with probability ranging from 0.03 to 0.06. Results are shown in Figure 13.

Whenever the target sequence appears in the action vocabulary, the model also correctly identifies it as causal.

The tendency to consistently segment the original actions versus the target part-action varied across parameter values, reflecting the tension between the causal and statistical segmentation cues in the corpus. Lower values of π and ω – representing a prior belief in causes being very unlikely, and relatively ineffectual – caused the model to overlook the causal information in favor of the statistical structure of the sequence (e.g., `TFD BL*R` or `TFD BL* R`), leading to actions being more likely, while higher values of π and ω made the causal cue more relevant, leading to the target part-action being more likely (e.g., `TFDBL* R`).

Summary. When presented with a corpus where statistical and causal cues to segmentation conflict, our joint inference model makes distinct segmentation and causal inference judgments, when compared to models that look only at transitional probabilities, or only at causal relationships. The joint model also outperforms these models in terms of successfully identifying both the target causal sequence, and the statistically determined actions used to create the corpus.

[6] For these parameters, the model always identifies the longer sequence, such as `TFDBL` as the causal sequence.

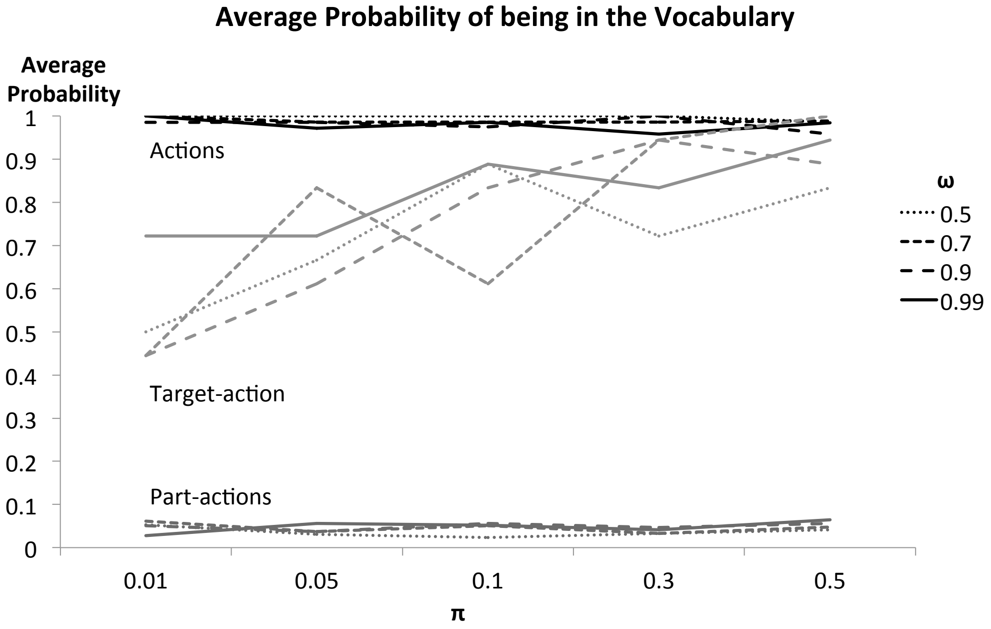


Figure 13. Average probability of a sequence appearing in the action vocabulary for Experiment 4 Joint Inference Model. Results are shown for a range of values of the two causal parameters, π and ω. Across all parameter values, actions are more likely than part-actions, with the target part-action being either intermediate, or equal to actions in probability.

Experiment 4: Conflicting Causal and Statistical Cues

Here we present people with the same corpora used to test our model and the alternative models described above, and again have them rate actions, part-actions, and non-actions on their coherence and causality. If participants preferentially use causal cues when judging causal relations, and statistical action structure when making segmentation judgments, then we would expect the target part-action to be rated as no more coherent than any other part-action (like the part-actions of Experiment 2) but to be rated as highly causal (like the target sequences of Experiments 3). Similarly, we would expect actions to be rated as highly coherent (like the actions in Experiment 2), but no more causal than the other non-target sequences (like the non-target sequences of Experiments 3).

However, if participants are using both sets of cues across inference tasks, then we would expect to see a compromise between the two in their ratings. In particular, we might expect that the target part-action would be rated not only as highly causal, but also as highly coherent (like the target sequences of Experiment 3), and similarly that actions might still be judged more causal than non-actions and part-actions (as in Experiment 2), even when the true causal sequence can be fully determined.

Method

Participants

Participants were 171 U.C. Berkeley undergraduates who received course credit for participating. 89 participants were assigned the causal condition and 82 participants were assigned to the coherence condition.

Stimuli. The structure and stimuli for this experiment closely matched those of Experiment 2. The second set of exposure corpora from Experiment 1 were used, with the same actions and non-actions as in Experiments 1 and 2, except that they were edited so that the target part-action in each corpus was always followed by a cartoon sound effect. In this experiment, the part-action `Empty Rattle Clean` was the target part-action in the first exposure corpus, and `Drink Blow Look` was the target part-action in the second corpus.

For the rating portion of this experiment, we created eight part-action comparison stimuli for each corpus – the four original part-action stimuli from Experiment 2, along with four additional part-actions. In both corpora, one of these part-actions was the target part-action. iMovie HD was used to assemble the test stimuli, as in the preceding experiments.

Procedure. Participants were randomly assigned to view one of the two exposure corpora, and were also randomly assigned to one of two follow-up question conditions. The instructions for the two conditions were identical to the causal condition and the coherence condition of Experiment 3a respectively. As in Experiments 3a and 3b, all participants were told that certain action sequences caused the bottle to play music.

Following the exposure corpus, participants in both conditions were presented with all 12 actions, non-actions, and part-actions individually, and asked to rate them by choosing a value on a 1 to 7 Likert scale. As in the previous experiments, in the *causal condition* participants were asked "How likely is this sequence to make the bottle play a musical sound?", with 1 representing "not likely" and 7 representing "most likely", while in the *coherence condition* participants were asked the question "How well does this action sequence go together?".

Results

Ratings from 86 participants in the *causal condition* were analyzed (data from an additional three participants who rated all sequences identically as either a 1 or 7 was discarded) using 2×4 ANOVAs on exposure corpus (1 or 2) and sequence type (action, non-action, part-action, target). No effects of exposure corpus were found. As predicted, there was an overall significant effect of sequence type $F(3, 252)= 111.04$ $MSE= 185.44$, $p < 0.0001$. The target part-action was rated as significantly more likely to cause a musical effect than actions $t(85)= -10.69$, $p < 0.0001$, one sample t-test on contrast values, part-actions $t(85)= -12.16$, $p < 0.0001$, one sample t-test on contrast values, and non-actions $t(85)= -11.33$, $p < 0.0001$, one sample t-test on contrast values.

Additionally, we looked at differences in ratings between the different types of non-target sequences. While all non-target sequences were rated as significantly less causal than the target sequence, actions were rated as significantly more likely to cause a musical effect than part-actions $t(85)= 4.29$, $p < 0.0001$, one sample t-test on contrast values, and than non-actions $t(85)= 2.93$, $p < 0.01$. There was no significant difference between ratings of part-actions and non-actions, $t(85)= 0.58$, $p = 0.57$, one sample t-test on contrast values.

Ratings from 82 participants in the *coherence condition* were analyzed (data from an additional two participants who rated all sequences identically as either a 1 or 7 was discarded) using 2×4 ANOVAs on exposure corpus (1 or 2) and sequence type (action,

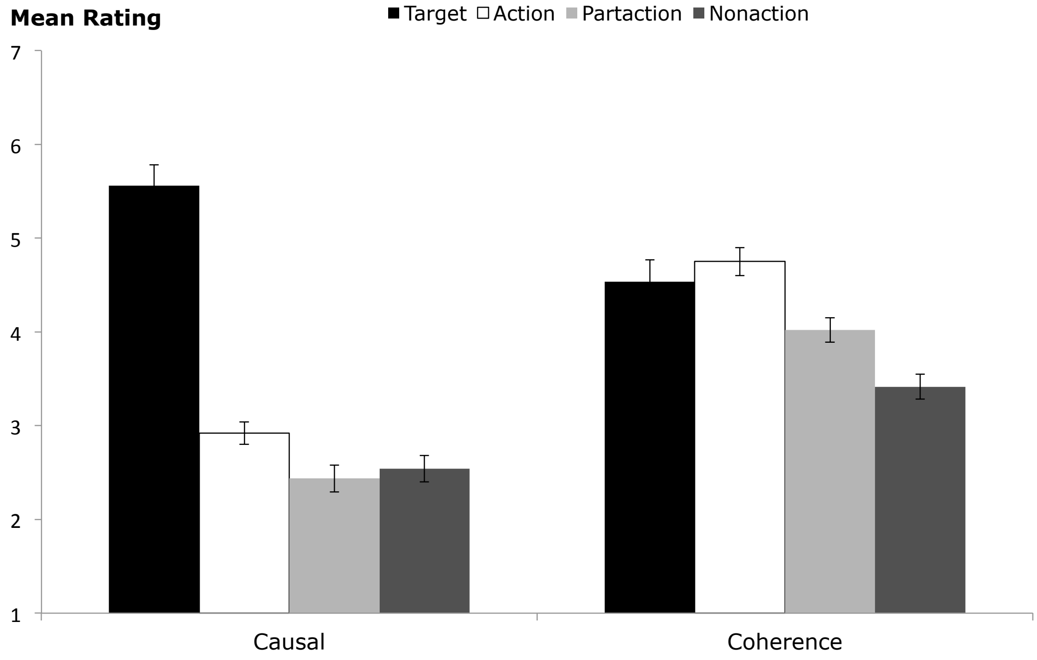


Figure 14. Results of Experiment 4. Error bars show one standard error.

non-action, part-action, target). No effects of exposure corpus were found. As predicted, there was an overall significant effect of sequence type $F(3, 234)= 17.39$, $MSE= 28.18$, $p < 0.0001$. Replicating the results of Experiment 2, actions were rated as going together significantly better than part-actions $t(79)= 7.39$, $p < 0.0001$, one sample t-test on contrast values, or non-actions $t(79)= 6.635$, $p < 0.0001$, one sample t-test on contrast values. There was also a significant difference between part-action and non-action ratings $t(79)= 3.97$, $p < 0.001$.

Additionally, the target part-action was rated as significantly more coherent than the non-target part-actions $t(79)= -2.24$, $p < 0.05$, and significantly more coherent than non-actions $t(79)= -4.16$, $p < 0.0001$. There was no significant difference in coherence ratings for the target part-action when compared to actions, $t(79)= 1.01$, $p = 0.32$, one sample t-test on contrast values.

Finally, we compared the two rating conditions directly, using a 2×4 ANOVAs on condition (causal or coherence) and sequence type (action, non-action, part-action, target). As expected, there were significant main effects of sequence type $F(3, 492)= 84.78$, $MSE= 139.49$, $p < 0.0001$, and condition, $F(1, 164)= 29.16$, $MSE= 108.90$, $p < 0.0001$, as well as a significant interaction between type and condition, $F(3, 492)= 41.60$, $MSE= 68.44$, $p < 0.0001$. Causal ratings were significantly higher for the target sequence than coherence ratings $t(164)= 3.17$, $p < 0.01$, while the reverse was true for actions, part-actions, and non-actions $t(164) \geq 4.50$, $p < 0.0001$.

Discussion

The results of Experiment 4 suggest that people take both causal relationships and statistical structure into account when interpreting continuous human behavior, correctly identifying the part-action as the most likely cause, but continuing to rate actions as more likely to also be causal when compared to other part-actions and non-actions. Similarly, they judged the causal part-action to be very cohesive, even though it violated the statistical regularities of the action sequence, suggesting that its causal properties led to it being considered a coherent unit of human action.

These results are consistent with our model results, and not consistent with the results of the alternative models we tested. In particular, if people are making separate causal and segmentation inferences, then the causal-only model predicts that the target part-action should be identified as causal, and that all other sequences should be rated equally non-causal, as under this model there is no more evidence for actions being causal than any other sequence type. Similarly, both our conditional probability and statistical action structure models would predict that the target part-action should not be rated as particularly coherent, since there is no more statistical evidence for it being in the action vocabulary than any other part-action sequence. Similarly, if people were not making a specifically causal inference, but were just treating the effect as another event in the sequence, they should not treat the target part-action as different from other part-actions in either condition.

General Discussion

In this work, we presented a Bayesian analysis of how statistical and causal cues to action segmentation should optimally be combined, as well as four experiments investigating human action segmentation and causal inference. We found that both adults and our model are sensitive to statistical regularities and causal structure in continuous action, and are able to combine these sources of information in order to correctly infer both causal relationships and segmentation boundaries. These results suggest that people jointly parse fluid motion into meaningful actions and infer the causal relevance of those actions, with both domains being learned simultaneously and mutually, rather than one being learned and then the other. In the following sections, we discuss the implications of these results, as well as limitations of the current work and future directions for research.

Implications

We used a non-parametric Bayesian model, adapted from work on statistical language processing to infer the segmentation and causal structure of the same sequences our human participants saw. The model represents our assumption that the same underlying process generates human actions and causal motion sequences, implicitly capturing that actions are being chosen intentionally, often to bring about causal outcomes. Our model results demonstrate that, at least in principle, action segmentation is learnable and may partly rely on domain general statistical learning mechanisms. The parallels in both human and computational model performance between word segmentation and action segmentation tasks similarly supports the possibility of a more general statistical learning ability at work in both domains. Our model also makes distinct segmentation and causal inference predictions when compared to models that look only at transitional probabilities, or only at causal relationships. Finally, these studies are the first to demonstrate that boundary judgments correspond to post-hoc sequence discrimination measures, suggesting that online segmentation and subsequent extraction of meaningful units arise from the same underlying processes.

Taken together, these four studies suggest that among the cues people use to segment action are both statistical cues such as transitional probabilities, and causal structure, and that action structure and causal structure are learned jointly rather than being layered one on top of the other. Adults, at least, can combine statistical regularities and causal structure to divide observed human behavior into meaningful actions. Adults can also use their inferred segmentation to help them identify likely causal actions. In particular, Experiments 3 and 4 demonstrate that people can identify the correct causal subsequence from within a longer set of fluid motion, a critical step in extracting higher-level goal directed units of behavior. In fact, these experiments are some of the first to demonstrate that people can carry out causal variable discovery within a continuous temporal stream of events. The fact that people rate artificially constructed actions as more coherent and meaningful than other motion sequences suggests that this is not an isolated statistical learning ability, but an integral part of action understanding. Finally, the results of Experiment 4 demonstrate that when statistical and causal cues are both present in the action stream, both of them influence people's judgments of action segmentation and of causal relationships.

We do not suggest that statistical cues are the only, or even the primary, way in which adults extract meaning from observed behavior (just as other cues are available in language, like prosody, grammatical structure, and of course meaning), but that it is *a* cue, and one that may be available, and especially important, early on. In particular, while goals and intentions may be important to extracting meaningful actions (e.g., Zacks, 2004; Zacks et al., 2009), they may often be opaque (e.g., to a young infant, when watching novel behavior). In these cases, statistical cues might be another option available to help break into the system.

As noted earlier in the paper, it has been theorized that statistical cues often arise naturally from goal-directed action. In fact, there has been some prior work showing that when higher level goals and intentions are removed from a video of everyday action (e.g., by playing the videos backwards Hard et al., 2006), adults are still able to segment these videos. In fact, work by Hard et al. (2006) suggests that their boundary judgments are relatively unaffected by their judgments of the intentionality of these sequences. Our studies provide evidence for a mechanism – tracking of low-level statistical patterns between motion elements – by which this segmentation could be accomplished.

Thus, an initially purely statistical parsing could lead to the discovery of intentionally meaningful segments, useful for developing an understanding of others' behavior in terms of goal-directed action. In fact, it has been hypothesized that just this type of statistical parsing could play an important role in social-learning in infants and non-human animals, allowing them to extract useful sequences to imitate, or to predict the actor's future behavior, without needing to fully understand their underlying intentions (Byrne, 1999).

There is of course evidence for other low level-cues aiding segmentation, such as changes in the overall image, or motion cues such as acceleration and deceleration of body parts, or changes in trajectory (e.g., Newtson et al., 1977; Reynolds et al., 2007; Hard et al., 2006, 2011). These accounts are not incompatible – image features and body pose may in fact be some of the features over which people track statistical relationships (e.g., Buchsbaum, Canini, & Griffiths, 2011; Reynolds et al., 2007).

However, there is reason to be particularly interested in the role of statistical information in particular, as there is a large body of evidence that infants (and adults) are able to pick up on just these sorts of distributional cues, and that this ability is extremely general and wide ranging (e.g., Saffran, Aslin, & Newport, 1996; Aslin et al., 1998; Saffran, Johnson, Aslin, & Newport, 1999; Kirkham et al., 2002; Fiser & Aslin, 2002). In fact, infants are not only capable of tracking statistical relationships over time-series such as actions and words, but even show some understanding of simultaneously observed population-level statistics over collections of objects (e.g., Xu & Garcia, 2008; Xu & Denison, 2009; Denison & Xu, 2010, 2014).

Similarly, the ability to track statistical relationships between variables in general, and conditional probability relationships in particular, lies at the heart of many accounts of human causal inference (e.g., Cheng, 1997; Shanks, 1995; Griffiths & Tenenbaum, 2005; Gopnik et al., 2004). That is, to reason about predictive relationships between objects in the environment in a causal manner, people must be able to detect those predictive relationships in the first place. There is evidence that at least this statistical aspect of causal reasoning begins emerging along roughly the same timeline as other statistical learning abilities (Sobel & Kirkham, 2006, 2007).

Therefore, both our modeling and experimental results provide support for the idea that this early-emerging, general statistical learning ability not only plays a role in extracting meaningful action units and identifying causal variables and causal relationships, but that it unites these two inference processes at a fundamental level. The qualitative fit of our joint inference model to human results, relative to alternative models, indicates that even without explicit higher-level goal structure present, inferences about causal relationships and action parsing do not proceed independently. This is consistent with evidence that infants and young children are biased to perceive causal events as inherently agentive (Muentener & Carey, 2010; Muentener, Bonawitz, Horowitz, & Schulz, 2012; Saxe, Tenenbaum, & Carey, 2005; Saxe, Tzelnic, & Carey, 2007; Bonawitz et al., 2010; Meltzoff, Waismeyer, & Gopnik, 2012), and may even assume that simple causal events (e.g., launching) are the result of agency, or an unseen human actor (Muentener & Carey, 2010; Saxe et al., 2005, 2007).

Limitations and Future Work

While these results are suggestive, there are number of outstanding questions and areas for future research.

First, as an ideal observer model, our model does not provide a process-level account of how people might be carrying out these inferences. In particular, our model uses a batch algorithm, meaning that the entire corpus is stored in memory, and then segmented as a whole. In contrast, people are able to provide boundary judgments in real time, as a sequence of actions is viewed (e.g., Newtson, 1973; Zacks, Tversky, & Iyer, 2001), and are of course subject to working memory constraints. In fact, a variety of work suggests that event segmentation both informs and results from working- and long-term memory representations (for a partial review see Kurby & Zacks, 2008). Frank et al. (2010) added memory constraints to the Goldwater et al. (2009) word segmentation model, and found that this increased the qualitative fit to human performance. A similar approach could be taken with our model in future work.

While the results of Experiment 1 suggest that statistical structure plays a role in online event perception, they are not definitive, since participants viewed an exposure corpus before providing online boundary judgments. Our current model is agnostic as to whether statistical cues are used immediately during the initial viewing, or only subsequently, to extract meaningful structure from what was observed. It is possible that our participants are first learning the actions (using statistics), and then using the recognized actions to find boundaries. This is an interesting mechanistic question beyond the scope of the current paper.

One possible mechanism for modeling process-level approximations to Bayesian inference is a particle filter (Doucet, de Freitas, & Gordon, 2001; Sanborn, Griffiths, & Navarro, 2010). Particle filters are a sequential algorithm for approximating a posterior distribution, using a discrete set of samples (*particles*) that are updated over time, as data is observed. They provide a natural mechanism for capturing both online processing constraints and working-memory limits in a Bayesian framework. This approach has been successfully applied to modeling process-level effects (e.g., causal order effects, garden path sentences) in human causal inference (Abbott & Griffiths, 2011; Levy, Reali, & Griffiths, 2009) and language processing, among other areas. A similar approach could be used to approximate our ideal observer model, and look for evidence of online versus post-hoc use on statistical segmentation by people.

This approach could also make contact with existing process-level accounts of event segmentation. For instance, Zacks and colleagues (e.g., Reynolds et al., 2007; Zacks, Speer, Swallow, Braver, & Reynolds, 2007; Zacks, Kurby, Eisenberg, & Haroutunian, 2011), argue that increases in prediction error correspond with boundary judgments in an action sequence, and trigger the segmentation and processing of the preceding event for long-term memory storage. Similarly, in a particle filter, large prediction errors can result in large weight changes or resampling of hypotheses, potentially providing a mechanism that is both consistent with our Bayesian computational-level account, and with an algorithmic-level account in which prediction error plays a significant role.

In addition, our current model used fixed parameter values, broadly chosen to match the structure of our stimuli. While a wide range of parameters values led to good model performance, equal to or superior to human performance (as we would expect for an ideal observer model), only a smaller subset led to perfect segmentation. It is currently an open question whether a similarly broad parameter range would lead to accurate segmentation of more naturalistic videos, or whether different types of action (e.g., a skilled surgeon conducting a specialist operation versus someone making a cup of coffee) would necessitate very different statistical assumptions. In addition, the model leaves open the question of how people might learn or calibrate their expectations about action sequences, and how much experience this learning process requires. While our current model used fixed parameter values, it could be extended to a *hierarchical* Bayesian model (e.g., Kemp, Perfors, & Tenenbaum, 2007; Heller, Sanborn, & Chater, 2009), where both the segmentation and segmentation parameters are simultaneously inferred.

Nonetheless, our current results provide us with reason to be optimistic that expectations about the statistical structure of a broad range of intentional actions might be learned quite quickly. Our participants were able to produce accurate boundary judgments and causal inferences after just 23 minutes of exposure, including just 90 viewings of each action, and 30 viewings of each part-action, despite the action combinations and statistical relationships in our corpus being deliberately arbitrary, so that only statistical cues in the corpus and not prior knowledge, could be used to segment the sequence.

By contrast, when viewing everyday intentional actions, even if the particular activity being viewed is novel (e.g., assembling a saxophone, as in Zacks, Tversky, & Iyer, 2001), adults have had years of experience observing similar intentional behaviors, during which they might learn more general statistical patterns (e.g., reaching precedes grasping), and even infants have likely had hundreds of hours of such observations. Given adults' success in our experiments after such a relatively short exposure, it is plausible to think that these patterns might be learned quite early on.

Our discussion of statistical cues has focused on sequential probabilities, rather than joint probabilities, of sequences. Experiments 3a and 3b demonstrated that people differentiate the true causal sequence from rearrangements of its component motions that appear with equal frequency, suggesting that the temporal order of these sequences is in fact important. Similarly, in Experiment 4, the causal part-action is rated as more coherent than the other part-actions, despite having the same joint probability.

In Experiments 1 and 2, we did not equate the joint probability of actions and part-actions, leaving open the possibility that, in these experiments, it was the joint rather than transitional probabilities of these sequences that people tracked. However, in work by Meyer and Baldwin (2011) and Stahl et al. (in press), the transitional (sequential) probabilities for actions were higher than for part-actions, while their joint probabilities were in some cases the same. These experiments found that both adults and infants remained able to discriminate actions from part-actions, though individual differences were evident in both cases. Exploring the extent to which people remain able to track these types of statistical patterns as the stimuli increase in complexity and variability remains an important question.

Like the computational model of word segmentation it extends, our model assumes that the lowest level of segmentation is already known (or pre-labeled). That is, that there is some sort of motion primitive (equivalent to a syllable in speech), that can already be recognized as a coherent unit. Since studies demonstrating human action segmentation have suggested that statistical patterns or features in human motion may correlate with segment boundaries at even the lowest level (e.g., Zacks, Tversky, & Iyer, 2001; Zacks, Braver, et al., 2001; Hard et al., 2006), we would like to see whether action boundaries can be automatically detected directly from video, without pre-existing knowledge of low-level motion units. A version of our model operating over low-level image features rather than discrete SMEs could help address this question.

Similarly, although the videos in the current studies featured a live actor carrying out natural object-directed motions, other aspects of the videos remain artificial by design – in order to focus on the statistical relationships between the small motion units, other cues such as motion changes (e.g., pauses, acceleration, deceleration), and the higher level goal structure of the actor were not present. Similarly, alternative cues to causality, such as observable physical or mechanical information or other perceptual cues (e.g., Michottean collision or launching events), were not included.

The actor was also observed in a somewhat simplified environment, interacting with only one object, which had just one causal property. Since we know that people can also successfully segment more naturalistic scenes (e.g., Zacks, Braver, et al., 2001; Zacks et al., 2010) with multiple objects, goals and sub-goals, and causal outcomes, one interesting direction to explore in future work is the extent to which joint statistical and causal inference contributes to our understanding of these more complex everyday scenes, and how low-level statistical information interacts with these other sources.

Just as it is important to explore how low-level motion cues might contribute to action segmentation in a bottom-up fashion, we would also like to understand how higher-level social information might contribute to action parsing. How might knowledge of an actor's goals and intentions influence segmentation? This is an especially interesting question in the context of hierarchical goal structures. Recent work suggests that people (e.g., Zacks & Tversky, 2001; Zacks, Tversky, & Iyer, 2001; Hard et al., 2006; Meyer, Baldwin, & Sage, 2011), and perhaps other apes (Byrne & Russon, 1998; Byrne, 1999, 2003 but also see Conway & Christiansen, 2001), naturally organize events into increasingly abstract hierarchical relationships, based on the underlying goals of the actors.

Once again, there are intuitive parallels to the language domain, where syllables are composed into words, which are in turn composed into phrases and sentences. An intriguing possibility is to see whether probabilistic models of phrase structure (e.g., Johnson, Griffiths, & Goldwater, 2007) could also be adapted to the action domain. Similarly, exploring whether there are garden path effects in action parsing akin to those that can occur in language (e.g., Levy, 2011) could help us better understand how action structure and goal structure are inferred.

Finally, while there are many obvious parallels between language processing and action processing, there are important differences as well. While it has sometimes been argued that language is a form of intentional action (Grice, 1957; Searle, 1969; Austin, 1962), the intent is fundamentally communicative, rather than instrumental, as in the case of much goal-directed action. It is not clear whether this leads to fundamentally different patterns of organization, especially at higher hierarchical levels. In addition to statistical structure, each domain has associated domain-specific knowledge and cues that play a large role in their interpretation, such as communicative pragmatics for language (Grice, 1957, 1989), and instrumental goals for action (e.g., Woodward, 1998; Gergely & Csibra, 2003; Sommerville, Woodward, & Needham, 2005).

In the context of our model, one important difference is that syllable meanings are tied solely to their segmentation into words – a sand witch and a sandwich have distinct non-overlapping meanings despite containing the identical syllables. By contrast, in some cases, the causal effects of motions do not have to be tied solely to the intentions underlying them – turning on the oven and then inserting a cake will cause the cake to bake, whether performed as a single coherent sequence or as two separate acts (or even by two separate actors). Our model does not currently capture this distinction, instead assuming that, just as the grouping of syllables can change their meaning, the grouping of motions can change their causal outcomes. It remains an open question to what extent this distinction influences action segmentation, both in terms of the natural statistics of action (people *could* perform causal motions separately from each other, but how often do they actually?), and in terms of people's causal inferences and assumptions when observing action.

Conclusion

In the real world, causal variables do not come pre-identified or occur in isolation, but instead are imbedded within a continuous temporal stream of events. Whether watching someone opening a door or making an object play music, a challenge faced by both human learners and machine learning algorithms is identifying subsequences that correspond to the appropriate variables for causal inference. Combining motion statistics with causal information may be one way for human (and non-human) learners to begin accomplishing this task.

Acknowledgements

We thank Meredith Meyer for discussions and feedback on this project, help with experimental design, and for providing experimental stimuli, and Jennifer Ng, Kimmy Yung, James Bach, Mia Krstic, Jonathan Lesser and Sophie Bridgers for help with data collection. This material is based upon work supported by a National Science Foundation Graduate Research Fellowship, the McDonnell Foundation Causal Learning Initiative, National Science Foundation Grant BCS-331620, and Grant FA9550-07-1-0351 from the Air Force Office of Scientific Research.

References

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the influence of particle filter parameters on order effects in causal learning. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471-517.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Austin, J. L. (1962). *How to do things with words.* Oxford University Press.

Baldwin, D. A., Andersson, A., Saffran, J. R., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382-1407.

Baldwin, D. A., Baird, J., Saylor, M., & Clark, A. (2001). Infants parse dynamic human action. *Child Development*, *72*(3), 708-717.

Bes, B., Sloman, S., Lucas, C. G., & Éric Raufaste. (2012). Non-bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*, 1178-1203.

Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., & Schulz, L. E. (2010). Just do it? investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, *115*, 104-107.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1-3), 71–105.

Buchsbaum, D., Canini, K. R., & Griffiths, T. L. (2011). Segmenting and recognizing human action using low-level video features. *Proc. of the 33rd Annual Conference of the Cognitive Science Society.*

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331-340.

Byrne, R. W. (1999). Imitation without intentionality. using string parsing to copy the organization of behaviour. *Animal Cognition*, *2*(2), 63-72.

Byrne, R. W. (2003). Imitation as behaviour parsing. *Philosophical Transactions: Biological Sciences*, *358*(1431), 529-536.

Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences*, *21*(5), 667-721.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(6), 367-405.

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*(12), 539-546.

Denison, S., & Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*, *13*, 798-803.

Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335-347.

Doucet, A., de Freitas, N., & Gordon, N. (2001). *Sequential Monte Carlo methods in practice.* New York: Springer.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate viual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429-433.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209-230.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences*, *99*(24), 15822-15826.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107-125.

Geisler, W. S. (2003). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (p. 825-838). MIT press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, *7*(7), 287–292.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Suffolk, UK: Chapman and Hall.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*, 39-60.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21-54.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178-186.

Goodman, N. D., Mansinghka, V. K., & Tenenbaum, J. B. (2007). Learning grounded causal models. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.*

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 1-31.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3), 254-260.

Grice, P. (1957). Meaning. *The Philosophical Review*, *67*, 377-388.

Grice, P. (1989). *Studies in the way of words.* Harvard University Press.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407-1455.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661–716.

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586-604.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory and Cognition*, *34*(6), 1221-1235.

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93-106.

Heller, K., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. *Advances in Neural Information Processing Systems*, 1-9.

Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology*, *45*(2), 575-585.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19.*

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307-321.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. *Cognition*, *83*(2), B35-B42.

Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671-680.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72-79.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.*.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. *Advances in Neural Information Processing Systems*, *21*.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323-1329.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 995-984.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*, 113-147.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: observational causal learning in 24-month-old infants. *Developmental Psychology*, *48*(5), 1215-1228.

Meyer, M., & Baldwin, D. A. (2011). Statistical learning of action: The role of conditional probability. *Learning and Behavior*, *39*(4), 383-398.

Meyer, M., Baldwin, D. A., & Sage, K. D. (2011). Assessing young children's hierarchical action segmentation. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.*

Meyer, M., DeCamp, P., Hard, B. M., & Baldwin, D. A. (2010). Assessing behavioral and computational approaches to naturalistic action segmentation. *Proc. of the 33nd Annual Conference of the Cognitive Science Society.*

Mirman, D., Magnuson, J. S., Graf Estes, K., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*(1), 271-280.

Muentener, P., Bonawitz, E., Horowitz, A., & Schulz, L. (2012). Mind the gap:

Investigating toddlers' sensitivity to contact relations in predictive events. *PLoS ONE*, *7*(4), 1-6.

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive Psychology*, *61*, 63-86.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*(1), 28-38.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847-862.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674-685.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1141-1159.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*, 613-643.

Roseberry, S., Richie, R., Hirsh-Pasek, K., Golinkoff, R. M., & Shipley, T. F. (2011). Babies catch a break : 7- to 9-month-olds track statistical probabilities in continuous dynamic events. *Psychological Science*, *22*(11), 1422-1424.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(27-52).

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606-621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101-105.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144-1167.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, *16*(12), 995-1001.

Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology*, *43*(1), 149-158.

Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition andDevelopment*, *8*(1), 113-128.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124-1139.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development*, *77*, 427-442.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language.* Cambrdige University Press.

Shanks, D. R. (1995). *The psychology of associative learning.* Cambridge University Press.

Sharon, T., & Wynn, K. (1998). Individuation of actions from continuous motion. *Psychological Science*, *9*(5), 357-362.

Sobel, D. M., & Kirkham, N. (2007). Bayes nets and babies: infants' developing statistical reasoning abilities and their representation of causal knowledge. *Developmental Science*, *10*(3), 298-306.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*(6), 1103-1115.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers.

Cognitive Science, *28*, 303-333.

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition*, *96*, B1-B11.

Speer, N. K., Swallow, K. M., & Zacks, J. M. (2003). Activation of human motion processing areas during event perception. *Cognitive, Affective, and Behavioral Neuroscience*, *3*(4), 335-345.

Stahl, A. E., Romberg, A. R., Roseberry, S., Golinkoff, R. M., & Hirsh-Pasek, K. (in press). Infants segment continuous events using transitional probabilities. *Child Development*.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems 13* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal induction. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (p. 35-42). Cambridge, MA: MIT Press.

Thimbleby, H. (2003). The directed chinese postman problem. *Software: Practice and Experience*, *33*(11), 1081–1096.

van Aardenne-Ehrenfest, T., & de Bruijn, N. G. (1951). Circuits and trees in oriented linear graphs. *Simon Stevin: Wis-en Natuurkundig Tijdschrift*, *28*, 203.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27*(3), 351-372.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523-541.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach infants selectively encode the goal object of an actor's reach. *Cognition*, *69*, 1-34.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, *11*(1), 73-77.

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and peopl. *Developmental Psychology*, *47*(5), 1220-1229.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97-104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.

Yeung, S., & Griffiths, T. L. (2011). Estimating human priors on causal strength. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*, 979-1008.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., … Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651-655.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201-216.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience*, *23*(12), 4057-4066.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind–brain perspective. *Psychological Bulletin*.

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: segmentation of narrative cinema. *Frontiers in Human Neuroscience*, *4*(1-15).

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3–21.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29–58.

Appendix A

Action Segmentation Model Details

We created a Bayesian rational learner model that jointly infers action segmentation and causal structure, using statistical regularities and temporal cues to causal relationships in an action stream. We use the nonparametric Bayesian model first used by Goldwater et al. (2009) as a model of action sequence generation, and also extend this model to incorporate causal information. Like the original word segmentation model, our model is based on a *Dirichlet process* (Ferguson, 1973), with actions composed of individual small motion elements taking the place of words composed of syllables. We model the generative process for creating a sequence of human actions as successively selecting actions to add to the stream, with the conditional probability of generating a particular action given by the *Chinese Restaurant Process* (Aldous, 1985), an easy to implement non-parametric process that is equivalent to the Dirichlet Process.

Generative Model for Action Sequences

An action sequence A is composed of a series of individual actions a_i which are in turn composed of individual motion units m_j. To create the sequence A we draw each a_i from G (a distribution of actions over all possible action sequences), where each action in G has an associated selection probability.

$$a_i|G \sim G \tag{9}$$

We in turn draw our distribution of actions G from a *Dirichlet Process* distribution, defined by the *concentration parameter* α_0 and the *base distribution* P_0.

$$G|\alpha_0, P_0 \sim DP(\alpha_0, P_0) \tag{10}$$

Here, P_0 is a distribution from which possible actions a_i are added to the vocabulary. In our model, the probability of including an item in the vocabulary is simply the product of the action's component motion unit probabilities, with an added assumption that action length is geometrically distributed (the longer the action, the less likely it is):

$$P_0(a_i = w) = p_{\#}(1 - p_{\#})^{n_w - 1} \prod_{j=1}^{n_w} p(m_{i,j}) \tag{11}$$

Where n_w is the length of $a_i = w$ in motion units, $p_{\#}$ and $(1 - p_{\#})$ are the probability of ending or not ending the action after each motion unit, and $p(m_{i,j})$ is the probability of the individual motion units that make up a_i. We assume a uniform probability over all motion units. Once an action a_i is drawn from P_0 it is added to G and assigned a probability in G determined by α_0.

We assume that like action length, action sequence length is also geometrically distributed:

$$P(A) = p_{\$}(1 - p_{\$})^{n-1} \prod_{i=1}^{n} p(a_i) \tag{12}$$

Where n is the length of A in actions, $p_{\$,}(1 - p_{\$})$ are the probability of ending or not ending the action sequence after a given action, and $p(a_i)$ is as described above.

Generative Model for Events

The action sequence A also contains effects e, which may occur both between and within actions. Some actions are causal actions, and are followed by effects with high probability. Each unique action type a_w has an associated binary variable $c_w \in \{0, 1\}$ that determines whether or not the action is causal:

$$c_w \sim Bernoulli(\pi_w) \tag{13}$$

Currently, we use a fixed value of π for all actions, but π_w may in turn be drawn from a *Beta* distribution in future versions of the model. If $c_w = 1$ then action w is causal, otherwise it is not. If an action is causal, then it is followed by an effect with probability ω_w. Again, we currently use a fixed value for ω. We use a small fixed value ϵ for the probability of an effect occurring anywhere in the sequence *other than* after a causal action.

Putting this all together, for each action a_i that is added to the sequence A (as described in the previous section), effects are or are not added after each of a_i's motion units with the following probabilities:

$$\begin{aligned} p(e|a_i = w, m_j, c_w = 1) &= \begin{cases} \omega_w, & j = n \\ \epsilon, & 0 \leq j < n \end{cases} \\ p(e|a_i = w, m_j, c_w = 0) &= \epsilon \end{aligned} \tag{14}$$

Where n_w is the length in motion units of $a_i = w$. In other words,the probability of inserting an effect after an internal motion unit is always a small constant (ϵ) across all actions, while the probability of inserting an effect at the end of an action is ϵ for non-causal actions and ω_w for causal actions.

Chinese Restaurant Process

Rather than explicitly drawing a vocabulary G from the Dirichlet Process, and then drawing the actions a_i from G in order to create the action sequence A, we would like to integrate across all possible vocabularies. This gives us the conditional probability of the next action in the sequence a_i, given all the previous actions $\mathbf{a}_{-i} = a_1 ... a_{i-1}$

$$p(a_i \mid \mathbf{a_{-i}}, \alpha_0, P_0) = \int p(a_i \mid G) p(G \mid \mathbf{a_{-i}}, \alpha_0, P_0) dG \tag{15}$$

It turns out that this conditional probability is equivalent to a simple construction known as the *Chinese Restaurant Process* (CRP). Here we use the CRP to formulate our generative model.

In the CRP customers enter a restaurant, and are seated at tables, each of which has an associated label. In this case, the associated labels represent actions. When the i^{th} customer enters the restaurant, the label at the table they sit at determines what the i^{th} action in our sequence will be. The probability of the i^{th} customer sitting at table $z_i = k$ is:

$$p(z_i = k|\mathbf{z}_{-i}) = \begin{cases} \frac{n_k}{n_{-i}+\alpha_0}, & 0 \leq k \leq K \\ \frac{\alpha_0}{n_{-i}+\alpha_0}, & k = K+1 \end{cases} \tag{16}$$

Where $n_{-i} = i - 1$ is the number of previously seated people, n_k is the number of customers already at table k, and K is the number of previously occupied tables. In other words, the probability of the i^{th} customer sitting at an already occupied table (i.e., choosing an action that has already appeared previously in the sequence A) depends on the proportion of customers already at that table, while the probability of them starting a new table depends on α_0.

Whenever a customer starts a new table, an action a_k must be associated with this table. This action is drawn from the distribution P_0, described above. Since multiple tables may be labeled with the same action, the probability that the next action in the sequence will have a particular value $a_i = w$ is:

$$p(a_i = w|\mathbf{a}_{-i}) = \frac{n_w}{n_{-i} + \alpha_0} + \frac{\alpha_0 P_0(a_i = w)}{n_{-i} + \alpha_0} \tag{17}$$

Where n_w is the number of times action w has appeared in the previous $\mathbf{a_{-i}}$ actions (the number of customers already seated at tables labeled with action w). In other words, the probability of a particular action $a_i = w$ being selected is based on the number of times it has already been selected (the probability of the i^{th} customer sitting at an existing table labeled with this action) and the probability of generating it anew (the probability of the customer sitting at a new table that is then assigned the label $a_k = w$).

Generative Process

We can now put together our complete generative model for creating a sequence of actions and effects. Our model parameters are sequence parameters $p_{\#}$, $p_{\$}$ and α_0, and causal parameters π, ω and ϵ.

1. Draw a probability distribution over actions, G

    (a) Actions are drawn from P_0

2. Select an action $a_i = w$ from G and add it to sequence A (this is equivalent to seating the i^{th} customer in the Chinese restaurant process)

    (a) Pick a table z_i for the i^{th} customer

        i. If it's a new table, draw a label from P_0

    (b) Add the label at table z_i to your list (in this case to the action sequence)

    (c) iterate until all customers are seated

3. Decide whether to insert any events after any of the motions $m_{i,j}$ composing $a_i = w$

    (a) For each of the $n_w - 1$ internal motion units in a_i insert an event with probability ϵ

    (b) If c_w is not yet known, draw c_w

    (c) Add an event after the last motion unit in $a_i = w$ with probability ω if $c_w = 1$ and with probability ϵ otherwise

4. With probability $p_{\$}$ repeat steps 2-4, otherwise terminate sequence A

Inference

Given an unsegmented action sequence, how do we find the boundaries between actions (find the correct segmented sequence)? For a given segmentation hypothesis h:

$$p(h|d) \propto p(d|h)p(h) \tag{18}$$

We want to infer the posterior distribution $p(h|d)$. A segmentation hypothesis h consists of whether or not there is an action boundary b after each motion m_j in the sequence. We can estimate $p(h|d)$ by iteratively considering one possible boundary at a time, while holding all other segment boundaries constant, a process known as Gibbs sampling. In deciding whether or not there should be an action boundary after motion m_j only two hypotheses need to be evaluated: $h_1 : b_j = false$ and $h_2 : b_j = true$. Since the segmentations defined by both $h_1$and h_2 will contain the same actions except for at the potential boundary point, only this difference in their probabilities needs to be considered. We'll call the segmentation boundaries that are the same in both hypotheses h^-. We will also refer to the single action generated under h_1 as w_1 and to the two actions generated under h_2 as w_2 and w_3. Going back to our generative model and the CRP, we can see that:

$$p(h_1|h^-, d) \propto p(w_1|h^-, d) = p(a_n = w_1|\mathbf{a_{-n}}) \tag{19}$$

where n is the total number of actions under h_1 and $\mathbf{a}_{-n} = a_1 ... a_{n-1}$ given by the segmentation h^-. this is because the CRP is *exchangeable*, which means we can treat w_1 as if it were the last action added to the sequence (the last person walking into the restaurant):

$$p(a_n = w_1|\mathbf{a}_{-n}) = \frac{n_{w1}}{n^- + \alpha_0} + \frac{\alpha_0 P_0(a_n = w_1)}{n^- + \alpha_0} = \frac{n_{w1} + \alpha_0 P_0(a_n = w_1)}{n^- + \alpha_0} \tag{20}$$

Where $n^- = n - 1$ is the number of actions under h^- (number of previously seated people) and n_{w1} is the number of times w_1 appears in h^- (the number of people already seated at tables labeled with w_1).

Similarly, the probability of h_2 is

$$\begin{aligned} p(h_2|h^-, d) &\propto p(w_2, w_3|h^-, d) \\ &= p(a_n = w_2|\mathbf{a_{-n}})p(a_{n+1} = w_3|a_n = w_2, \mathbf{a_{-n}}) \end{aligned} \tag{21}$$

As in h_1, the probability of w_2 is simply:

$$p(a_n = w_2|\mathbf{a}_{-n}) = \frac{n_{w1}}{n^- + \alpha_0} + \frac{\alpha_0 P_0(a_n = w_2)}{n^- + \alpha_0} = \frac{n_{w2} + \alpha_0 P_0(a_n = w_2)}{n^- + \alpha_0} \tag{22}$$

The probability of w_3 is a little different, since it depends on w_2 as well as h^-

$$\begin{aligned} p(a_{n+1} = w_3|\mathbf{a}_{-n}, a_n = w_2) &= \frac{n_{w3} + I(w_2 = w_3)}{n^- + 1 + \alpha_0} + \frac{\alpha_0 P_0(a_{n+1} = w_3)}{n^- + 1 + \alpha_0} \\ &= \frac{n_{w3} + I(w_2 = w_3) + \alpha_0 P_0(a_{n+1} = w_3)}{n + \alpha_0} \end{aligned} \tag{23}$$

Where n_{w3} is the number of times w_3 appears in h^- and $I(w_2 = w_3) = 1$ if w_2 and w_3 are the same (in other words $n_{w3} + I(w_2 = w_3)$ is the number of customers already seated at tables labeled with w_3). Also, $n^- + 1 = n$ is the number of actions in $h^- + w_2$ (the number of previously seated customers, before w_3).

Finally, since h_2 hypothesizes an action sequence one longer than h_1 we need to consider the probability of having a sequence of this increased length. Putting all of this together

$$\begin{aligned} p(h_2|h^-, d) &\propto p(length = n + 1, w_2, w_3|h^-, d) \\ &= p(length = n + 1|h^-, d)p(w_2|h^-, d)p(w_3|w_2, h^-, d) \\ &= (1 - p_\$) \cdot \frac{n_{w2} + \alpha_0 P_0(a_n = w_2)}{n^- + \alpha_0} \cdot \frac{n_{w3} + I(w_2 = w_3) + \alpha_0 P_0(a_{n+1} = w_3)}{n + \alpha_0} \end{aligned}$$

Using Information from Events in the Action Sequence

If the action sequence we're trying to segment also contains events, we can use this information to aid our segmentation. In particular, whether or not there is an event e_j at the segmentation boundary being considered, or an event e_k following the action created by the segmentation, impacts our probability estimates for h_1 and h_2:

$$p(h_1|h^-, d) \propto p(a_n = w_1|\mathbf{a}_{-n}) \cdot p(e_j|a_n = w_1, c_{w1}, \mathbf{c_{-n}}, \mathbf{e}_{h^-}) \cdot p(e_k|a_n = w_1, c_{w1}, \mathbf{c_{-n}}, \mathbf{e}_{h^-}) \quad (24)$$

Where $\mathbf{e}_{h^-}$ are all the events that occured in h^-, c_{w1} is the causal variable for $a_n = w_1$ and $\mathbf{c_{-n}}$ are the causal variables for the other actions in the sequence. Since h_1 predicts no boundary at position j:

$$p(e_j|\cdot) = p(e_j|a_n = w_1) = \begin{cases} \epsilon, & e_j = 1 \\ 1-\epsilon, & e_j = 0 \end{cases} \quad (25)$$

The computation for e_k is slightly more complicated. Assuming we know the value of c_{w1} (we'll discuss sampling the $\mathbf{c}$ values below) then:

$$\begin{aligned} p(e_k = 1|\cdot) = p(e_j = 1|a_n = w_1, c_{w1}) &= \begin{cases} \epsilon, & c_{w1} = 0 \\ \omega, & c_{w1} = 1 \end{cases} \\ p(e_k = 0|\cdot) = p(e_j = 1|a_n = w_1, c_{w1}) &= \begin{cases} 1-\epsilon, & c_{w1} = 0 \\ 1-\omega, & c_{w1} = 1 \end{cases} \end{aligned} \quad (26)$$

If we haven't yet sampled a value for c_{w1} then we simply sum across possible cases, in which case:

$$p(e_k|\cdot) = \begin{cases} \epsilon + \omega, & e_k = 1 \\ (1-\epsilon) + (1-\omega), & e_k = 0 \end{cases} \quad (27)$$

We perform similar computations for h_2, except that in this case, both e_j and e_k occur at the ends of actions (and so are computed like e_k above):

$$
\begin{aligned}
p(e_j = 1|\cdot) = p(e_j = 1|a_n = w_2, c_{w2}) &= \begin{cases} \epsilon, & c_{w2} = 0 \\ \omega, & c_{w2} = 1 \end{cases} \\
p(e_j = 0|\cdot) = p(e_j = 1|a_n = w_2, c_{w2}) &= \begin{cases} 1 - \epsilon, & c_{w2} = 0 \\ 1 - \omega, & c_{w2} = 1 \end{cases} \\
p(e_k = 1|\cdot) = p(e_k = 1|a_{n+1} = w_3, c_{w3}) &= \begin{cases} \epsilon, & c_{w3} = 0 \\ \omega, & c_{w3} = 1 \end{cases} \\
p(e_k = 0|\cdot) = p(e_k = 1|a_{n+1} = w_3, c_{w3}) &= \begin{cases} 1 - \epsilon, & c_{w3} = 0 \\ 1 - \omega, & c_{w3} = 1 \end{cases}
\end{aligned}
\tag{28}
$$

The probabilities when c_{w2} and c_{w3} have not yet been sampled are identical to when c_{w1} has not yet been sampled.

Inferring Which Actions are Causal

Just as we can use the causal variables $\mathbf{c}$ to help infer the action segmentation, we can use the action segmentation to help infer the causal variable values. In this case our hypothesis h consists of values c_w for all actions. Just as we can estimate the action segmentation by iteratively considering one boundary at a time, we can estimate the values for $\mathbf{c}$ by looking at one action a_w at a time, and estimating c_w, while holding the remaining c values constant. In this case h_1 is the hypothesis that $c_w = 1$ (a_w is causal) and h_2 is the hypothesis that c_w= 0 (a_w is not causal).

$$
p(h_1|h^-, d) \propto p(c_w = 1|\mathbf{a}_n, \mathbf{e}_n) = p(c_w = 1|\mathbf{a_w},\mathbf{e_w}) \tag{29}
$$

$$
p(h_2|h^-, d) \propto p(c_w = 0|\mathbf{a}_n, \mathbf{e}_n) = p(c_w = 0|\mathbf{a_w},\mathbf{e_w}) \tag{30}
$$

Where $\mathbf{a_w}$ is all occurrences of action w in the sequence, and $\mathbf{e_w}$ is all effects (or lack thereof) following these occurrences. In this case

$$p(c_w|\mathbf{a}_n, \mathbf{e}_n) \propto p(\mathbf{e_n}|\mathbf{a}_n, c_w) \cdot p(c_w) = p(c_w) \prod_{i=0}^{n} p(e_i|c_w, a_i) \tag{31}$$

With

$$p(c_w) = \begin{cases} \pi, & c_w = 1 \\ 1 - \pi & c_w = 0 \end{cases}$$

$$p(e_i = 1|c_w, a_i) = \begin{cases} \omega, & c_w = 1 \\ \epsilon, & c_w = 0 \end{cases} \tag{32}$$

$$p(e_i = 0|c_w, a_i) = \begin{cases} 1 - \omega, & c_w = 1 \\ 1 - \epsilon, & c_w = 0 \end{cases}$$

Putting this all together

$$\begin{aligned} p(h_1|h^-, d) &\propto p(c_w = 1|\mathbf{a}_n, \mathbf{e}_n) \propto \pi \cdot \omega^{ne_w^+} \cdot (1 - w)^{ne_w^-} \\ p(h_2|h^-, d) &\propto p(c_w = 0|\mathbf{a}_n, \mathbf{e}_n) \propto (1 - \pi) \cdot \epsilon^{ne_w^+} \cdot (1 - \epsilon)^{ne_w^-} \end{aligned} \tag{33}$$

Where ne_w^+ is the number of times action w is followed by an event and ne_w^- is the number of times it's not followed by an event.

Gibbs Sampling

To summarize, our algorithm for discovering the best segmentation of an unsegmented action sequence is:

1. For each motion unit m_j in the action sequence
2. Decide whether there should be a boundary after this motion unit
   (a) Hold the rest of the segmentation constant

(b) Calculate the probability of $h_1 =$ no boundary and $h_2 =$ boundary

(c) Normalize probabilities and decide probabilistically between h_1 and h_2

3. iterate 1 and 2 until the segmentation converges and/or a pre-determined stopping point is reached.

Simulated Annealing. The Gibbs sampling procedure described above has a number of advantages. It is relatively simple to implement, and once the sampler converges it produces samples from the true posterior distribution. However, changes are made locally, one boundary at a time. Searching through the hypothesis space may therefore require passing through many low probability segmentations in order to reach higher probability hypotheses, causing convergence to be slow.

To address this issue, we used an approach known as *simulated annealing* (Kirkpatrick, Gelatt, & Vecchi, 1983). This approach broadens exploration of the hypothesis space early on in sampling by making the relative probabilities of the different hypotheses more uniform. Using the metaphor of “slow cooling”, annealing uses a *temperature* parameter γ to gradually adjust the probability of moving to a particular hypothesis. γ starts at a high value, and is slowly reduced to 1 over the course of sampling.

Using simulated annealing, we sample our boundary probabilities using $p(h_1|h^-, d)^{\frac{1}{\gamma}}$ and $p(h_2|h^-, d)^{\frac{1}{\gamma}}$. Notice that when $\gamma = 1$ this is just regular Gibbs sampling. When $\gamma > 1$ the relative probabilities of the two hypotheses are more uniform, making transitions to lower probability segmentations more likely.

Following Goldwater et al. (2009), for our simulations, we ran each sampler for 20,000 iterations, annealing in 10 increments of 2000 iterations each, with $\frac{1}{\gamma} = (.1, .2, ..., .9, 1)$. For each simulation, we ran three randomly seeded samplers, each initialized from a random segmentation of the input corpus, and averaged results from 10 samples drawn from the last 1,000 iterations of each sampler, to estimate the posterior distributions and evaluate the model. This allowed for an additional burn-in period of 1000 samples with $\gamma = 1$.

Appendix B

De Bruijn Sequence

For an alphabet A of size k, a *De Bruijn sequence* $B(k, n)$ is a cyclical sequence within which each subsequence of length n appears exactly once as a consecutive sequence. The sequence is constructed by first creating a *De Bruijn graph*, where every sequence of size $n - 1$ appears as a node, and outgoing edges represent a sequence of n items – the $n - 1$ items of the node the edge is leaving, and the item labeling the edge itself. See Figure B1 for an example graph. The sequence is then created by traversing the graph in a *Eulerian cycle* – a path through the graph that traverses each edge exactly once.

To create the exposure corpora for Experiment 3, we used a $B(4, 3)$ De Bruijn sequence (creating length three sequences from a length four alphabet), modifying the process slightly, by only allowing edges in the graph that would not cause the resulting sequence of three items to contain a repeated item (in other words, each two-item node has exactly two outgoing edges). There are a number of algorithms for finding the shortest path through a graph that traverses each edge at least once (which will always be the Eulerian cycle if it exists). In this work we used the approach presented by Thimbleby (2003).

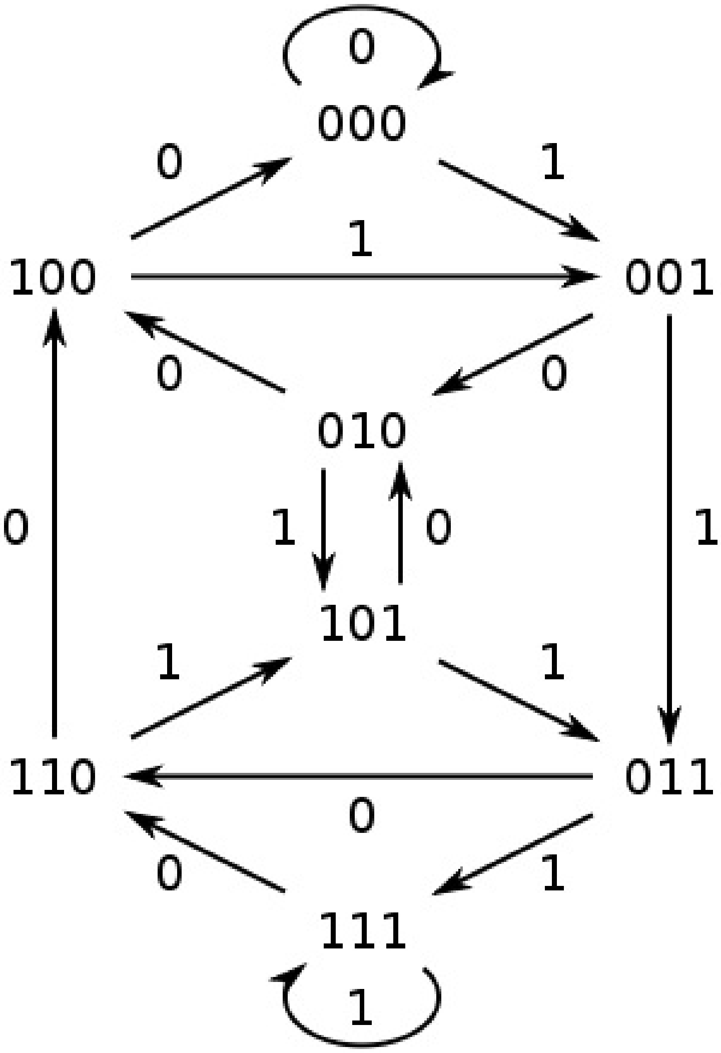


Figure B1. Example De Bruijn Graph (image created by Michael Hardy). If you traverse the graph in a Eulerian cycle, passing through every edge exactly once, then every four-digit sequence occurs exactly once, creating a $B(2, 4)$ De Bruijn sequence

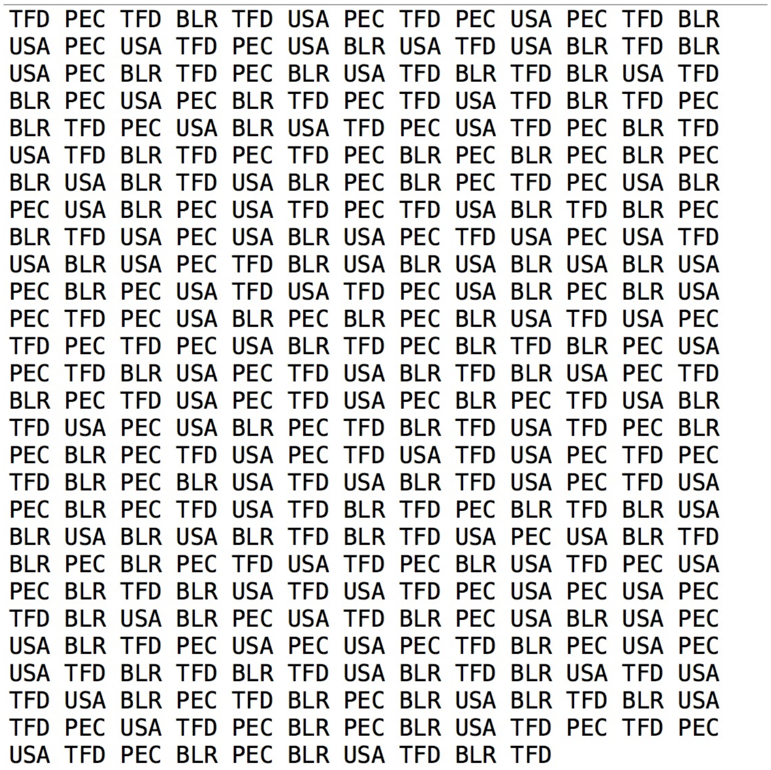
Appendix C

Example Corpora

Here we show some examples of the unsegmented corpora used as input for our computational simulations, as well as examples of potential segmentations. Figure C1, shows an example of one of the unsegmented corpora used in Experiments 1 and 2, while Figure C2 shows the correct segmentation for this corpus, which puts boundaries between all of the actions, and nowhere else. Figures C4 and C5 similarly show examples of an unsegmented and segmented corpus for Experiments 3a and 3b. Recall that in these experiments there was causal structure present in the corpus, but no statistical structure, so that a "correct" segmentation identifies all occurrences of the causal triplet (with additional boundaries being considered neither correct nor incorrect). Finally, Figure C6 shows an example of an unsegmented corpus used in Experiment 4. Notice that it is the same as the example corpus from Experiment 1, except that an effect follows each occurrence of the part-action `TFDBL`. Figure C7 shows an example of a segmentation that compromises between the causal and statistical cues in this corpus, identifying the causal part-action as well as the actions used to create the corpus.

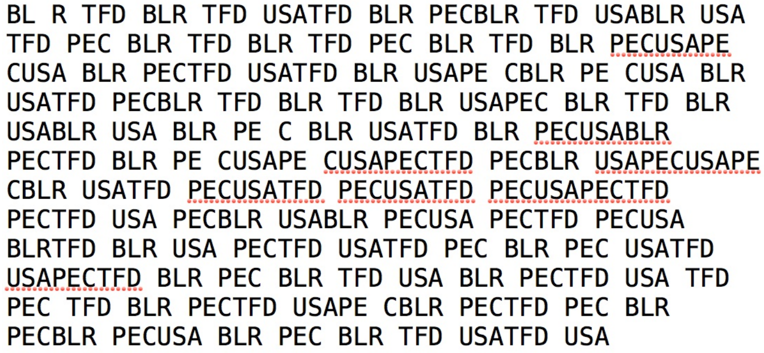
```
TFDPECTFDBLRTFDUSAPECTFDPECUSAPECTFDBLRUSAPECUSATFDPECU
SABLRUSATFDUSABLRTFDBLRUSAPECBLRTFDPECBLRUSATFDBLRTFDBL
RUSATFDBLRPECUSAPECBLRTFDPECTFDUSATFDBLRTFDPECBLRTFDPEC
USABLRUSATFDPECUSATFDPECBLRTFDUSATFDBLRTFDPECTFDPECBLRP
ECBLRPECBLRPECBLRUSABLRTFDUSABLRPECBLRPECTFDPECUSABLRPE
CUSABLRPECUSATFDPECTFDUSABLRTFDBLRPECBLRTFDUSAPECUSABLR
USAPECTFDUSAPECUSATFDUSABLRUSAPECTFDBLRUSABLRUSABLRUSAB
LRUSAPECBLRPECUSATFDUSATFDPECUSABLRPECBLRUSAPECTFDPECUS
ABLRPECBLRPECBLRUSATFDUSAPECTFDPECTFDPECUSABLRTFDPECBLR
TFDBLRPECUSAPECTFDBLRUSAPECTFDUSABLRTFDBLRUSAPECTFDBLRP
ECTFDUSAPECTFDUSAPECBLRPECTFDUSABLRTFDUSAPECUSABLRPECTF
DBLRTFDUSATFDPECBLRPECBLRPECTFDUSAPECTFDUSATFDUSAPECTFD
PECTFDBLRPECBLRUSATFDUSABLRTFDUSAPECTFDUSAPECBLRPECTFDU
SATFDBLRTFDPECBLRTFDBLRUSABLRUSABLRUSABLRTFDBLRTFDUSAPE
CUSABLRTFDBLRPECBLRPECTFDUSATFDPECBLRUSATFDPECUSAPECBLR
TFDBLRUSATFDUSATFDPECUSAPECUSAPECTFDBLRUSABLRPECUSATFDB
LRPECUSABLRUSAPECUSABLRTFDPECUSAPECUSAPECTFDBLRPECUSAPE
CUSATFDBLRTFDBLRTFDUSABLRTFDBLRUSATFDUSATFDUSABLRPECTFD
BLRPECBLRUSABLRTFDBLRUSATFDPECUSATFDPECBLRPECBLRUSATFDP
ECTFDPECUSATFDPECBLRPECBLRUSATFDBLRTFD
```

Figure C1. Example unsegmented corpus for Experiments 1 and 2 (line breaks are for display, and were not present in the input)



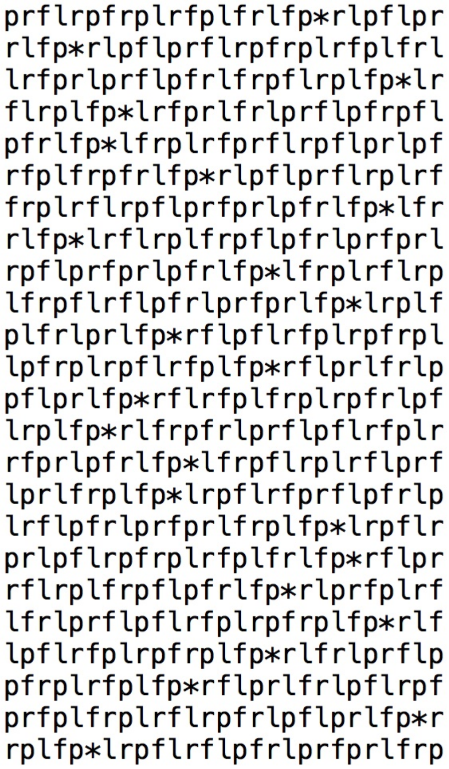
```
TFD PEC TFD BLR TFD USA PEC TFD PEC USA PEC TFD BLR
USA PEC USA TFD PEC USA BLR USA TFD USA BLR TFD BLR
USA PEC BLR TFD PEC BLR USA TFD BLR TFD BLR USA TFD
BLR PEC USA PEC BLR TFD PEC TFD USA TFD BLR TFD PEC
BLR TFD PEC USA BLR USA TFD PEC USA TFD PEC BLR TFD
USA TFD BLR TFD PEC TFD PEC BLR PEC BLR PEC BLR PEC
BLR USA BLR TFD USA BLR PEC BLR PEC TFD PEC USA BLR
PEC USA BLR PEC USA TFD PEC TFD USA BLR TFD BLR PEC
BLR TFD USA PEC USA BLR USA PEC TFD USA PEC USA TFD
USA BLR USA PEC TFD BLR USA BLR USA BLR USA BLR USA
PEC BLR PEC USA TFD USA TFD PEC USA BLR PEC BLR USA
PEC TFD PEC USA BLR PEC BLR PEC BLR USA TFD USA PEC
TFD PEC TFD PEC USA BLR TFD PEC BLR TFD BLR PEC USA
PEC TFD BLR USA PEC TFD USA BLR TFD BLR USA PEC TFD
BLR PEC TFD USA PEC TFD USA PEC BLR PEC TFD USA BLR
TFD USA PEC USA BLR PEC TFD BLR TFD USA TFD PEC BLR
PEC BLR PEC TFD USA PEC TFD USA TFD USA PEC TFD PEC
TFD BLR PEC BLR USA TFD USA BLR TFD USA PEC TFD USA
PEC BLR PEC TFD USA TFD BLR TFD PEC BLR TFD BLR USA
BLR USA BLR USA BLR TFD BLR TFD USA PEC USA BLR TFD
BLR PEC BLR PEC TFD USA TFD PEC BLR USA TFD PEC USA
PEC BLR TFD BLR USA TFD USA TFD PEC USA PEC USA PEC
TFD BLR USA BLR PEC USA TFD BLR PEC USA BLR USA PEC
USA BLR TFD PEC USA PEC USA PEC TFD BLR PEC USA PEC
USA TFD BLR TFD BLR TFD USA BLR TFD BLR USA TFD USA
TFD USA BLR PEC TFD BLR PEC BLR USA BLR TFD BLR USA
TFD PEC USA TFD PEC BLR PEC BLR USA TFD PEC TFD PEC
USA TFD PEC BLR PEC BLR USA TFD BLR TFD
```

Figure C2. Example true segmentation for Experiments 1 and 2



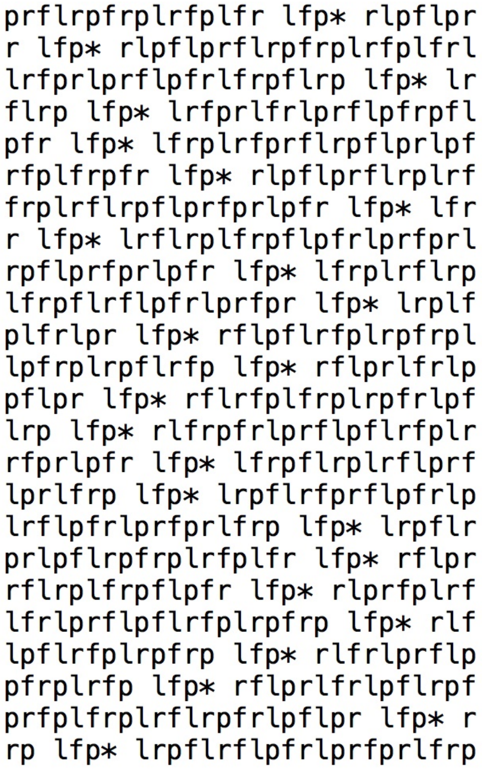
```
BL R TFD BLR TFD USATFD BLR PECBLR TFD USABLR USA
TFD PEC BLR TFD BLR TFD PEC BLR TFD BLR PECUSAPE
CUSA BLR PECTFD USATFD BLR USAPE CBLR PE CUSA BLR
USATFD PECBLR TFD BLR TFD BLR USAPEC BLR TFD BLR
USABLR USA BLR PE C BLR USATFD BLR PECUSABLR
PECTFD BLR PE CUSAPE CUSAPECTFD PECBLR USAPECUSAPE
CBLR USATFD PECUSATFD PECUSATFD PECUSAPECTFD
PECTFD USA PECBLR USABLR PECUSA PECTFD PECUSA
BLRTFD BLR USA PECTFD USATFD PEC BLR PEC USATFD
USAPECTFD BLR PEC BLR TFD USA BLR PECTFD USA TFD
PEC TFD BLR PECTFD USAPE CBLR PECTFD PEC BLR
PECBLR PECUSA BLR PEC BLR TFD USATFD USA
```

Figure C3. Group segmentation produced by above-chance participants in Experiment 1b



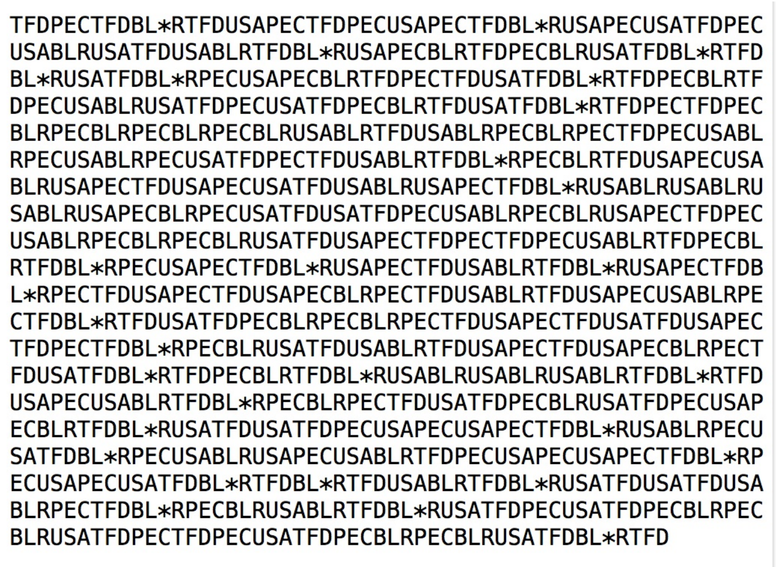
```
prflrpfrplrfplfrlfp*rlpflpr
rlfp*rlpflprflrpfrplrfplfrl
lrfprlprflpfrlfrpflrplfp*lr
flrplfp*lrfprlfrlprflpfrpfl
pfrlfp*lfrplrfprflrpflprlpf
rfplfrpfrlfp*rlpflprflrplrf
frplrflrpflprfprlpfrlfp*lfr
rlfp*lrflrplfrpflpfrlprfprl
rpflprfprlpfrlfp*lfrplrflrp
lfrpflrflpfrlprfprlfp*lrplf
plfrlprlfp*rflpflrfplrpfrpl
lpfrplrpflrfplfp*rflprlfrlp
pflprlfp*rflrfplfrplrpfrlpf
lrplfp*rlfrpfrlprflpflrfplr
rfprlpfrlfp*lfrpflrplrflprf
lprlfrplfp*lrpflrfprflpfrlp
lrflpfrlprfprlfrplfp*lrpflr
prlpflrpfrplrfplfrlfp*rflpr
rflrplfrpflpfrlfp*rlprfplrf
lfrlprflpflrfplrpfrplfp*rlf
lpflrfplrpfrplfp*rlfrlprflp
pfrplrfplfp*rflprlfrlpflrpf
prfplfrplrflrpfrlpflprlfp*r
rplfp*lrpflrflpfrlprfprlfrp
```

Figure C4. Example unsegmented corpus for Experiments 3a and 3b



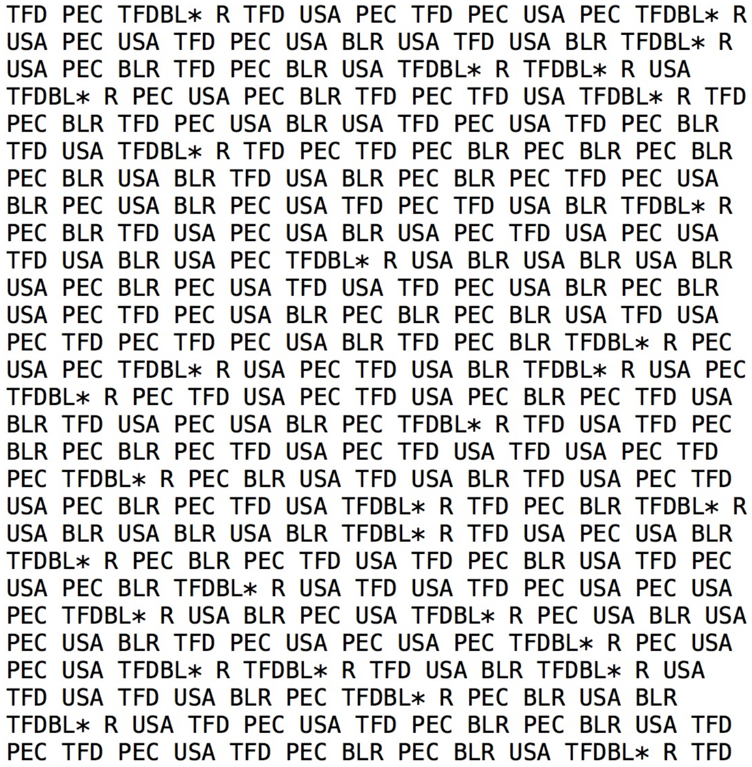
```
prflrpfrplrfplfr lfp* rlpflpr
r lfp* rlpflprflrpfrplrfplfrl
lrfprlprflpfrlfrpflrp lfp* lr
flrp lfp* lrfprlfrlprflpfrpfl
pfr lfp* lfrplrfprflrpflprlpf
rfplfrpfr lfp* rlpflprflrplrf
frplrflrpflprfprlpfr lfp* lfr
r lfp* lrflrplfrpflpfrlprfprl
rpflprfprlpfr lfp* lfrplrflrp
lfrpflrflpfrlprfpr lfp* lrplf
plfrlpr lfp* rflpflrfplrpfrpl
lpfrplrpflrfp lfp* rflprlfrlp
pflpr lfp* rflrfplfrplrpfrlpf
lrp lfp* rlfrpfrlprflpflrfplr
rfprlpfr lfp* lfrpflrplrflprf
lprlfrp lfp* lrpflrfprflpfrlp
lrflpfrlprfprlfrp lfp* lrpflr
prlpflrpfrplrfplfr lfp* rflpr
rflrplfrpflpfr lfp* rlprfplrf
lfrlprflpflrfplrpfrp lfp* rlf
lpflrfplrpfrp lfp* rlfrlprflp
pfrplrfp lfp* rflprlfrlpflrpf
prfplfrplrflrpfrlpflpr lfp* r
rp lfp* lrpflrflpfrlprfprlfrp
```

Figure C5. Example true segmentation for Experiments 3a and 3b



```
TFDPECTFDBL*RTFDUSAPECTFDPECUSAPECTFDBL*RUSAPECUSATFDPEC
USABLRUSATFDUSABLRTFDBL*RUSAPECBLRTFDPECBLRUSATFDBL*RTFD
BL*RUSATFDBL*RPECUSAPECBLRTFDPECTFDUSATFDBL*RTFDPECBLRTF
DPECUSABLRUSATFDPECUSATFDPECBLRTFDUSATFDBL*RTFDPECTFDPEC
BLRPECBLRPECBLRPECBLRUSABLRTFDUSABLRPECBLRPECTFDPECUSABL
RPECUSABLRPECUSATFDPECTFDUSABLRTFDBL*RPECBLRTFDUSAPECUSA
BLRUSAPECTFDUSAPECUSATFDUSABLRUSAPECTFDBL*RUSABLRUSABLRU
SABLRUSAPECBLRPECUSATFDUSATFDPECUSABLRPECBLRUSAPECTFDPEC
USABLRPECBLRPECBLRUSATFDUSAPECTFDPECTFDPECUSABLRTFDPECBL
RTFDBL*RPECUSAPECTFDBL*RUSAPECTFDUSABLRTFDBL*RUSAPECTFDB
L*RPECTFDUSAPECTFDUSAPECBLRPECTFDUSABLRTFDUSAPECUSABLRPE
CTFDBL*RTFDUSATFDPECBLRPECBLRPECTFDUSAPECTFDUSATFDUSAPEC
TFDPECTFDBL*RPECBLRUSATFDUSABLRTFDUSAPECTFDUSAPECBLRPECT
FDUSATFDBL*RTFDPECBLRTFDBL*RUSABLRUSABLRUSABLRTFDBL*RTFD
USAPECUSABLRTFDBL*RPECBLRPECTFDUSATFDPECBLRUSATFDPECUSAP
ECBLRTFDBL*RUSATFDUSATFDPECUSAPECUSAPECTFDBL*RUSABLRPECU
SATFDBL*RPECUSABLRUSAPECUSABLRTFDPECUSAPECUSAPECTFDBL*RP
ECUSAPECUSATFDBL*RTFDBL*RTFDUSABLRTFDBL*RUSATFDUSATFDUSA
BLRPECTFDBL*RPECBLRUSABLRTFDBL*RUSATFDPECUSATFDPECBLRPEC
BLRUSATFDPECTFDPECUSATFDPECBLRPECBLRUSATFDBL*RTFD
```

Figure C6. Example unsegmented corpus for Experiment 4 (line breaks are for display, and were not present in the input)



```
TFD PEC TFDBL* R TFD USA PEC TFD PEC USA PEC TFDBL* R
USA PEC USA TFD PEC USA BLR USA TFD USA BLR TFDBL* R
USA PEC BLR TFD PEC BLR USA TFDBL* R TFDBL* R USA
TFDBL* R PEC USA PEC BLR TFD PEC TFD USA TFDBL* R TFD
PEC BLR TFD PEC USA BLR USA TFD PEC USA TFD PEC BLR
TFD USA TFDBL* R TFD PEC TFD PEC BLR PEC BLR PEC BLR
PEC BLR USA BLR TFD USA BLR PEC BLR PEC TFD PEC USA
BLR PEC USA BLR PEC USA TFD PEC TFD USA BLR TFDBL* R
PEC BLR TFD USA PEC USA BLR USA PEC TFD USA PEC USA
TFD USA BLR USA PEC TFDBL* R USA BLR USA BLR USA BLR
USA PEC BLR PEC USA TFD USA TFD PEC USA BLR PEC BLR
USA PEC TFD PEC USA BLR PEC BLR PEC BLR USA TFD USA
PEC TFD PEC TFD PEC USA BLR TFD PEC BLR TFDBL* R PEC
USA PEC TFDBL* R USA PEC TFD USA BLR TFDBL* R USA PEC
TFDBL* R PEC TFD USA PEC TFD USA PEC BLR PEC TFD USA
BLR TFD USA PEC USA BLR PEC TFDBL* R TFD USA TFD PEC
BLR PEC BLR PEC TFD USA PEC TFD USA TFD USA PEC TFD
PEC TFDBL* R PEC BLR USA TFD USA BLR TFD USA PEC TFD
USA PEC BLR PEC TFD USA TFDBL* R TFD PEC BLR TFDBL* R
USA BLR USA BLR USA BLR TFDBL* R TFD USA PEC USA BLR
TFDBL* R PEC BLR PEC TFD USA TFD PEC BLR USA TFD PEC
USA PEC BLR TFDBL* R USA TFD USA TFD PEC USA PEC USA
PEC TFDBL* R USA BLR PEC USA TFDBL* R PEC USA BLR USA
PEC USA BLR TFD PEC USA PEC USA PEC TFDBL* R PEC USA
PEC USA TFDBL* R TFDBL* R TFD USA BLR TFDBL* R USA
TFD USA TFD USA BLR PEC TFDBL* R PEC BLR USA BLR
TFDBL* R USA TFD PEC USA TFD PEC BLR PEC BLR USA TFD
PEC TFD PEC USA TFD PEC BLR PEC BLR USA TFDBL* R TFD
```

Figure C7. Example compromise segmentation for Experiment 4